

Aplicação de Modelos de Aprendizado de Máquina na Predição da Desistência Acadêmica em Cursos de TI do IFPE

Verlaynne Kelley da Hora Rocha Araujo
Instituto Federal de Pernambuco
Palmares, Brasil
verlaynne.araujo@palmares.ifpe.edu.br

Yanka Mirelly Ayala de Sales Silva
Instituto Federal de Pernambuco
Palmares, Brasil
ymass@discente.ifpe.edu.br

Maria Fernanda De Lima Oliveira Sales
Instituto Federal de Pernambuco
Palmares, Brasil
mflos@discente.ifpe.edu.br

Diogo Lopes da Silva
Instituto Federal de Pernambuco
Palmares, Brasil
diogo.lopes@palmares.ifpe.edu.br

Abstract—Dropout rates are one of the main challenges faced by educational institutions, especially in the Exact Sciences, which have high dropout rates. This study, applied to the IFPE context, proposes the use of machine learning techniques to predict student dropout rates based on academic, demographic, and socioeconomic data from students between the 2001–2025 academic years. After data preprocessing, which included variable coding, missing value processing, and class balancing with SMOTE, three supervised models were trained: Decision Tree, Extra Tree and Multilayer Perceptron (MLP). The models were adjusted using Grid Search and evaluated using metrics such as accuracy, precision, recall, and F1-score. The MLP performed best, standing out for its ability to more accurately identify students prone to dropout. The results demonstrate the potential of artificial intelligence as a tool to support educational management, favoring the adoption of more effective preventive measures.

Keywords—Artificial neural networks. Decision trees. Extra trees. Academic dropout. Machine learning. IFPE.

Resumo—A evasão acadêmica constitui um dos principais desafios enfrentados pelas instituições de ensino, especialmente nos cursos de Ciências Exatas, que apresentam altos índices de desistência. Este estudo, aplicado ao contexto do IFPE, propõe a utilização de técnicas de aprendizado de máquina para prever a evasão estudantil, com base em dados acadêmicos, demográficos e socioeconômicos de alunos entre os períodos de 2001.1 e 2025.1. Após o pré-processamento dos dados, que incluiu codificação de variáveis, tratamento de valores ausentes e balanceamento das classes com SMOTE, foram treinados três modelos supervisionados: Decision Tree, Extra Tree e Multilayer Perceptron (MLP). Os modelos foram ajustados via Grid Search e avaliados por métricas como acurácia, precisão, recall e F1-score. O MLP apresentou o melhor desempenho, destacando-se pela capacidade de identificar com maior precisão os alunos propensos à evasão. Os resultados demonstram o potencial da inteligência artificial como ferramenta

de apoio à gestão educacional, favorecendo a adoção de medidas preventivas mais eficazes.

Palavras-chave—Redes neurais artificiais. Decision Tree. Extra Trees. Evasão acadêmica. Aprendizado de máquina. IFPE.

I. INTRODUÇÃO

A evasão acadêmica é um problema complexo que impacta significativamente as instituições de ensino, especialmente em cursos de exatas, estendido pela complexidade das disciplinas que resultam em desafios aos estudantes. No contexto do Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE), o problema também é um desafio, afetando todas as modalidades de ensino (Integrado, Subsequente e Superior), com reflexos diretos no desempenho institucional e no futuro dos alunos [1].

A evasão não é apenas um indicador de insucesso acadêmico, mas também um reflexo de questões socioeconômicas, psicológicas e pedagógicas que permeiam a vida dos estudantes [2]. De acordo com [3], fatores como dificuldades financeiras, necessidade de conciliar trabalho e estudo, baixa adaptação ao curso e à instituição, além de deficiências na formação básica, estão entre as principais causas desse fenômeno. No ensino técnico federal, como o oferecido pelo IFPE, a evasão afeta diretamente as políticas públicas de inclusão e formação profissional. Segundo [3], cursos da área de exatas apresentam maiores índices de desistência, o que reforça a necessidade de ações preventivas fundamentadas em dados concretos. Diante desse cenário, a identificação precoce de alunos em risco de evasão torna-se uma estratégia importante para a implementação de ações preventivas e personalizadas, visando

à permanência e à conclusão bem-sucedida do curso. Pesquisas recentes demonstram que técnicas de aprendizado de máquina vêm sendo aplicadas com sucesso em diferentes contextos preditivos, inclusive na identificação de estudantes em risco de evasão [4], [5]. Este trabalho tem como objetivo principal aplicar três abordagens de aprendizado de máquina: a rede neural MLP (Multilayer Perceptron), o algoritmo de árvore de decisão Decision Tree Classifier e o Extra Trees Classifier, para prever, de forma individualizada, a evasão acadêmica nos cursos de TI do IFPE. Utilizando dados acadêmicos e demográficos dos estudantes, busca-se explorar o potencial desses modelos preditivos na tarefa de classificação da evasão, comparando seus desempenhos a fim de identificar quais técnicas oferecem maior precisão e aplicabilidade no apoio à gestão educacional e à tomada de decisões institucionais.

II. FUNDAMENTAÇÃO TEÓRICA

A seguir, são estabelecidas as bases conceituais necessárias para a compreensão da metodologia utilizada. O objetivo é explorar as técnicas de Aprendizado de Máquina aplicadas à predição da evasão escolar, com foco nos três algoritmos utilizados: Decision Tree, Extra Trees e Multilayer Perceptron. A aplicação de modelos computacionais para análise de dados educacionais tem se mostrado uma área de grande relevância. Um estudo relacionado foi realizado por [4], no qual os autores propuseram um método baseado em redes neurais do tipo Fuzzy ARTMAP para prever o risco de evasão escolar. O trabalho apresentou resultados promissores na identificação de grupos de risco, reforçando a eficácia de abordagens baseadas em aprendizado de máquina para lidar com situações de evasão no contexto educacional brasileiro.

A. Aprendizado de Máquina

O Aprendizado de Máquina (Machine Learning) é um subcampo da Inteligência Artificial dedicado ao desenvolvimento de algoritmos que permitem aos computadores aprender e tomar decisões a partir de dados, sem serem programados para cada tarefa. O processo consiste em treinar um modelo com um conjunto de dados, permitindo que ele identifique padrões, realize classificações ou faça previsões. As abordagens de aprendizado são comumente divididas em supervisionadas, não supervisionadas e por reforço. Este estudo se concentra no Aprendizado Supervisionado, no qual o modelo é treinado com um conjunto de dados rotulados, ou seja, cada entrada de dados possui uma saída ou resposta correta [6]. O objetivo é que o modelo aprenda a mapear as entradas para as saídas correspondentes, tornando-se capaz de prever o resultado para novos dados nunca antes vistos, mas seguindo as mesmas variáveis. A predição da evasão escolar é um problema de

classificação supervisionada, onde o objetivo é classificar um aluno como “evadido” ou “não evadido”.

B. Rede Neural Artificial

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados na estrutura e no funcionamento do cérebro humano. Elas são compostas por unidades de processamento interconectadas, chamadas de neurônios artificiais, organizadas em camadas. A estrutura mais comum inclui uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada conexão entre neurônios possui um peso associado, que é ajustado durante o processo de treinamento. O modelo aprende” ao otimizar esses pesos para minimizar a diferença entre suas previsões e os resultados reais, um processo geralmente conduzido pelo algoritmo de backpropagation. Devido à sua capacidade de modelar relações complexas e não lineares, as RNAs são amplamente utilizadas em problemas de alta complexidade.

C. Classificador Extra Trees

O algoritmo Extra Trees Classifier (Extremely Randomized Trees) é uma técnica de aprendizado de máquina do tipo ensemble, que constroi múltiplas árvores de decisão a partir de amostras aleatórias dos dados e utiliza critérios aleatórios para as divisões dos nós. Diferentemente de outros métodos baseados em árvores, o Extra Trees seleciona aleatoriamente os pontos de divisão, o que aumenta a diversidade entre as árvores geradas. Essa aleatoriedade contribui para a redução da variância do modelo, tornando-o mais robusto contra o problema de overfitting, que ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, perdendo sua capacidade de generalização. A aplicação do Extra Trees no domínio educacional justifica-se por sua alta capacidade de generalização e robustez. Ao introduzir aleatoriedade nos pontos de divisão dos nós, o algoritmo constroi um modelo ensemble menos suscetível à variância e ao ruído, características comuns em dados socioeconômicos e de desempenho estudantil. Dessa forma, se torna altamente eficaz em tarefas de predição e apoio à decisão, como a identificação de alunos propensos à evasão, a análise de desempenho e a identificação de perfis para o planejamento institucional. Para esta pesquisa, foi utilizada a implementação do algoritmo Extra Trees disponibilizada pela biblioteca Scikit-learn, amplamente reconhecida por sua eficiência e facilidade de uso em aplicações de aprendizado supervisionado.

D. Classificador Multilayer Perceptron

Optamos por utilizar a rede neural do tipo MLP por sua facilidade de aplicação e sua boa performance com dados não lineares, característica importante ao lidarmos com uma base que contém informações socioeconômicas e demográficas. Para

que o modelo fosse capaz de processar corretamente essas variáveis, aplicamos a técnica One Hot Encoding, conforme citado anteriormente, convertendo variáveis categóricas em formato numérico. Esse algoritmo é inspirado no funcionamento das redes neurais biológicas, cuja base teórica foi inicialmente proposta por [7], ao descreverem um modelo lógico para representar a atividade neural por meio de neurônios artificiais. O MLP é composto por múltiplas camadas de neurônios interconectados e, durante o treinamento, utiliza a técnica de backpropagation, na qual os erros obtidos na saída são retropropagados para ajustar os pesos das conexões sinápticas, promovendo melhorias gradativas no desempenho do modelo. Dessa forma, a escolha pelo MLP se justifica pela sua versatilidade, pela capacidade de generalização e pelo seu bom desempenho em tarefas supervisionadas com variáveis de alta dimensionalidade e natureza heterogênea, características observadas na base de dados utilizada neste estudo.

III. TRABALHOS RELACIONADOS

Um estudo relacionado foi realizado por [8], no qual os autores propuseram um método baseado em redes neurais do tipo Fuzzy ARTMAP para prever o risco de evasão escolar. O trabalho apresentou resultados promissores na identificação de grupos de risco, reforçando a eficácia de abordagens baseadas em aprendizado de máquina para o enfrentamento da evasão no contexto educacional brasileiro. Outro trabalho relevante foi desenvolvido por [9], que aplicou modelos de redes neurais do tipo RBF e MLP para analisar a evasão no curso de Sistemas de Informação da UFRJ. A RBF (Radial Basis Function) é uma rede neural que utiliza funções de ativação baseadas na distância entre os dados e pontos centrais, sendo eficaz na identificação de padrões locais. A autora explorou diferentes variáveis acadêmicas e demográficas, demonstrando o potencial dessas técnicas na identificação de perfis com maior probabilidade de evasão. Um estudo realizado por [3] analisou as razões do abandono escolar entre jovens no Brasil, destacando fatores sociais, econômicos e institucionais que contribuem para a evasão. Embora não tenha utilizado técnicas de aprendizado de máquina, a pesquisa fornece uma base importante para compreender o fenômeno da evasão sob uma perspectiva sociológica, o que complementa abordagens computacionais como as utilizadas neste trabalho. Este trabalho se diferencia por focar na construção de modelos preditivos voltados à identificação individualizada de estudantes com risco de evasão nos cursos de exatas do IFPE, utilizando dados institucionais e demográficos. Ao contrário de abordagens estatísticas tradicionais, a proposta é oferecer uma aplicação prática que auxilie diretamente a gestão acadêmica, possibilitando intervenções mais rápidas e eficazes. Embora a técnica utilizada se assemelhe aos critérios considerados no cálculo da PNP como a análise

de fatores acadêmicos e demográficos, neste trabalho deu-se ênfase à predição do risco de evasão com base em modelos de aprendizado de máquina, e não ao cálculo direto da evasão em um período específico. Isso reforça o caráter antecipatório da proposta, voltado à tomada de decisão preventiva.

IV. METODOLOGIA

Para o desenvolvimento deste trabalho, diversos procedimentos foram realizados, abrangendo desde a extração dos dados até o treinamento e a comparação dos modelos preditivos. Na Figura 1, é apresentada uma visão geral da metodologia seguida neste estudo.

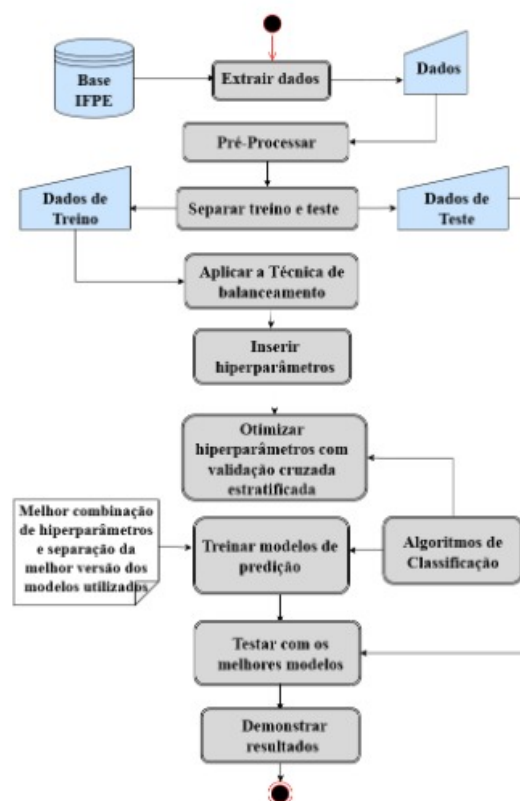


Figura 1: Etapas da metodologia aplicada.

A. Conjunto de Dados

Os dados utilizados neste trabalho foram fornecidos, de forma anonimizada, pelo Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE), abrangendo registros do período entre o semestre 2001.1 e 2025.1, exclusivamente de cursos da área de exatas. Inicialmente, foi realizada uma análise exploratória da base com o objetivo de compreender e estabelecer relações entre as colunas disponíveis. O conjunto de

dados, no formato de CSV, continha 50 colunas e 16.264 linhas. Para fins de rotulagem, foram considerados alunos evadidos aqueles com situação de matrícula em uma das seguintes categorias: Abandono/Evasão, Cancelamento Voluntário, Jubilado, Cancelamento Compulsório ou Transferido Externo. Já os alunos não evadidos foram identificados pelas categorias: Formado, Concluído e Aguardando Colação de Grau. Além disso, registros vinculados a situações como Matrícula Vínculo Institucional, Matriculado, Aguardando ENADE, Concludente, Trancado, Estagiário (Concludente), Transferido Interno, Falecido e Projeto Final (Concludente) foram removidos da base, por representarem alunos com matrícula ainda ativa ou em transição. Como o modelo desenvolvido é supervisionado, é necessário trabalhar apenas com dados rotulados, ou seja, com alunos que já possuem um desfecho conhecido, evasão ou não evasão. No total, a base final contou com 8.727 alunos evadidos e 2.904 alunos não evadidos, totalizando 11.631 registros.

B. Técnicas de Pré-processamento e Balanceamento de Dados

Inicialmente, foram removidas colunas que apresentavam identificadores únicos, informações redundantes, alta taxa de valores ausentes ou que poderiam causar vazamento de dados. As colunas Cod_Aluno, cod_matricula e Matrícula foram excluídas por conterem identificadores únicos que não contribuem para a análise preditiva.

Colunas como Cod_instituicao, Cod_curso, Turma Atual, Cod_cidade e Nível/Regime de Ensino foram consideradas irrelevantes, pois seus dados já estavam representados em outras colunas com informações equivalentes. A coluna Desc_Matriz_Curricular também foi excluída por sua baixa relevância. Outras colunas foram removidas devido à elevada porcentagem de dados faltantes, como Pólo Municipal (90,76%), necessidades_especiais (98,51%), N_Membros_Familia (88,79%), Dt_Conclusao_Curso (80,10%), Renda Familiar Per Capita PNP (63,42%), Renda Familiar Per Capita SIG (83%), Percentual Frequência (81,77%), Area Procedência Escola Origem (52%) e Grupo Étnico, cuja ausência de dados era praticamente total (99,98%).

A coluna Renda Familiar também foi removida, pois além de apresentar 63,19% de dados ausentes, sua informação já está contemplada pela coluna Renda Familiar Per Capita, que foi mantida no conjunto de dados e fornece uma medida mais normalizada para fins de modelagem. Também foram excluídas colunas que poderiam induzir o modelo a um aprendizado incorreto ou enviesado por vazamento de dados, foram elas: Último Evento de Matrícula, Sit. Últ. Per. Letivo, Última presença e Ano_Conclusao_Ensino_Medio, por conterem informações que não estariam disponíveis no momento da previsão.

Além disso, a coluna EIXO_TECNOLÓGICO foi descartada por não oferecer variabilidade útil, visto que todos os cursos

pertenciam à área de tecnologia. Por fim, a coluna Situação Matrícula foi retirada do conjunto de entrada por se tratar da variável alvo da análise. Ao todo, 25 colunas foram excluídas. Após essa limpeza e a adição da variável alvo, a base de dados foi consolidada com 11.631 linhas e as 26 colunas, apresentadas na Tabela I.

Nome da Coluna	Descrição
Desc_Instituicao	Nome da instituição e do campus.
DESCRICAO_CURSO	Nome do curso
Nivel_Ensino	Nível acadêmico do curso
Qtd Períodos	Duração total do curso em períodos/semestres.
Turno Ingresso	Turno em que o aluno ingressou no curso.
Turno	Turno atual do aluno.
Período Atual	Período/semestre que o aluno está cursando.
Per. Let. Ingresso	Ano e semestre de entrada do aluno.
Regime Internato	Indica se o aluno está em regime de internato.
Desc_Modalidade_Curso	Modalidade específica do curso (Tecnólogo, Subsequente).
Sexo	Gênero do aluno.
Nascimento	Data de nascimento do aluno.
Naturalidade	Cidade e estado de nascimento do aluno.
Cor/Raça	Autodeclaração de cor ou raça do aluno.
Desc_Forma_Ingresso_Matricula	Método de ingresso do aluno na instituição.
Cota	Indica se o aluno ingressou por meio de cotas.
Renda Familiar Per Capita	Faixa de renda por pessoa na família do aluno.
Escola de Origem	Tipo de escola anterior do aluno (Pública, Privada).
C.E.P.	Código de Endereçamento Postal do aluno.
Cidade	Cidade de residência do aluno.
Bairro	Bairro de residência do aluno.
Coefficiente_Rendimento	Média de notas/desempenho acadêmico do aluno.
Desc_Estado_Civil	Estado civil do aluno.
Cod_Nacionalidade	Código da nacionalidade do aluno.
MODALIDADE_ENSINO	Formato do curso (Presencial ou EAD)
Evasão	Indicador de abandono do curso (1 para sim, 0 para não).

Tabela I: Descrição das variáveis da base de dados final.

Nossa base contempla dados dos seguintes cursos: Análise e Desenvolvimento de Sistemas, Técnico em Manutenção e Suporte em Informática, Técnico em Informática para Internet,

Operador de Computador, Engenharia de Software, Técnico em Informática, Licenciatura em Computação, Técnico em Desenvolvimento de Sistemas, Sistemas para Internet, Técnico em Redes de Computadores e Técnico em Computação Gráfica. Como mencionado anteriormente, foi realizada a rotulagem dos casos de evasão e não evasão. Após essa etapa e a exclusão das colunas desnecessárias, foi possível visualizar o gráfico de evasão, Figura 2, contendo alunos evadidos e não evadidos, respectivamente: : 8727 e 2904.

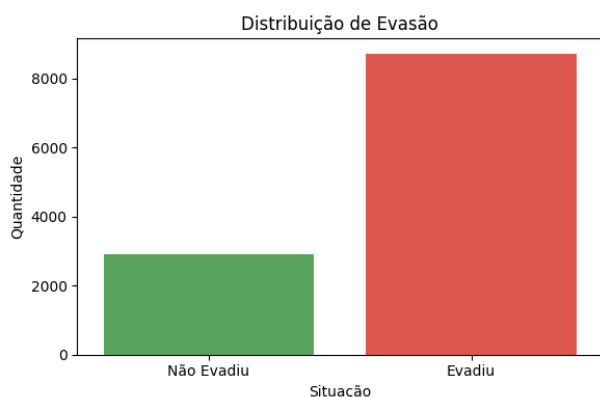


Figura 2: Gráfico da distribuição de evasão a partir dos dados extraídos.

Em seguida, as colunas foram separadas entre categóricas e numéricas, a fim de aplicar técnicas específicas de pré-processamento. Para as colunas numéricas com valores ausentes, utilizou-se a mediana como forma de imputação. Já nas colunas categóricas com dados faltantes, os valores foram substituídos por "Desconhecido", e foi aplicada a técnica de codificação one-hot encoding. A base de dados foi inicialmente dividida em 80% para treinamento e 20% para teste, respeitando a proporcionalidade das classes por meio da técnica de stratified sampling. Considerando o desbalanceamento presente entre as classes com um número significativamente menor de registros da classe representando a evasão, foi aplicada a técnica SMOTE (Synthetic Minority Over-sampling Technique) no conjunto de treinamento. O SMOTE atua gerando novas amostras sintéticas da classe minoritária com base na interpolação entre exemplos reais vizinhos, evitando a simples duplicação de registros. Essa abordagem busca melhorar a capacidade de generalização dos modelos, prevenindo o viés em favor da classe majoritária e possibilitando que o algoritmo aprenda padrões representativos de ambas as classes de forma equilibrada. Os hiperparâmetros foram configurados de forma a influenciar diretamente a capacidade de generalização dos modelos. Para realizar esse ajuste de maneira criteriosa, utilizou-se a técnica de validação cruzada estratificada combinada com

a busca em grade (Grid Search). Nessa abordagem, diversas combinações de configurações dos modelos são testadas e avaliadas por meio de subdivisões do conjunto de dados de treinamento. A validação cruzada estratificada garante que a distribuição das classes seja preservada em cada subdivisão, assegurando uma avaliação mais justa e representativa.

C. Métricas de Avaliação

Nesse estudo foram utilizadas métricas de avaliação do desempenho dos modelos preditivos utilizados. Essas métricas possibilitam uma análise detalhada da qualidade das previsões, especialmente em contextos com classes desbalanceadas, como é o caso da evasão acadêmica [10]. A seguir, os conceitos dos elementos:

Verdadeiro Positivo (TP): quando o modelo identifica corretamente um estudante que evadiu. Verdadeiro Negativo (TN): quando o modelo acertou ao prever que o estudante permaneceu. Falso Positivo (FP): quando o modelo prevê evasão, mas o estudante permanece. Falso Negativo (FN): quando o modelo não detecta a evasão, mas o estudante efetivamente desiste.

A partir dessas definições, é possível derivar indicadores estatísticos fundamentais para a avaliação do desempenho dos modelos preditivos. As métricas foram:

Acurácia: expressa a proporção total de acertos, considerando tanto as predições corretas de evasão quanto de permanência, em relação ao total de casos avaliados. Precisão: quantifica a proporção de acertos entre todas as predições positivas realizadas, indicando o quanto as previsões de evasão são confiáveis. Recall: mensura a capacidade do modelo em identificar corretamente os estudantes que realmente evadiram, sendo crucial em contextos onde se busca minimizar falsos negativos. F1-Score: Corresponde à média harmônica entre precisão e recall, funcionando como uma métrica balanceada para cenários em que há desequilíbrio entre as classes, como é comum em bases com baixa taxa de evasão. Essas métricas oferecem uma visão abrangente do desempenho dos modelos, equilibrando a taxa de acertos, a confiabilidade das predições e a detecção efetiva dos estudantes em risco [10].

D. Comparativo dos Modelos

Conforme citado nas subseções C, D e E, os classificadores Extra Trees, Decision Tree e MLP possuem características distintas que os tornam adequados para a tarefa de predição da evasão acadêmica, a depender do contexto e dos objetivos da aplicação. O Extra Trees Classifier, abordado na subseção C, mantém a base interpretável dos modelos de árvore, mas se diferencia por incorporar aleatoriedade tanto na seleção das amostras quanto nas divisões dos nós. Essa característica torna o modelo mais robusto, favorecendo uma melhor capacidade de generalização, o que o torna uma opção interessante quando

se busca equilíbrio entre interpretabilidade e desempenho. O Decision Tree Classifier, descrito na subseção D, é reconhecido por sua simplicidade e interpretabilidade. Sua estrutura em forma de árvore facilita a visualização das regras de decisão, o que permite compreender com clareza os fatores que influenciam a evasão. Por essa razão, é especialmente útil em cenários onde a transparência do modelo é fundamental. No entanto, sua estrutura única o torna mais suscetível ao superajuste, especialmente em bases complexas. O MLP, descrito na subseção E, é uma rede neural de múltiplas camadas capaz de modelar relações não lineares entre variáveis. Sua aplicação é especialmente recomendada em conjuntos de dados com alta dimensionalidade e natureza heterogênea, como ocorre em bases educacionais que envolvem informações acadêmicas, demográficas e socioeconômicas. Apesar da menor interpretabilidade em relação aos modelos baseados em árvores, o MLP oferece maior flexibilidade na modelagem de padrões complexos. Cada técnica apresenta vantagens específicas, por isso, a escolha entre elas é levada em consideração o equilíbrio desejado entre transparência, capacidade de generalização e complexidade dos dados.

V. RESULTADOS

Após a realização do treinamento e ajuste dos modelos por meio da técnica de validação cruzada e a utilização do Grid Search, foram obtidos os valores das principais métricas de avaliação: acurácia, precisão, recall e F1-score. Essas métricas foram calculadas com base nos dados de teste, previamente separados, e têm como objetivo mensurar o desempenho dos algoritmos na tarefa de prever a evasão acadêmica. Esses resultados reforçam a importância de alinhar a escolha dos modelos às características da base de dados e aos objetivos da tarefa. Em cenários como o da evasão acadêmica, com dados heterogêneos e complexos, modelos capazes de capturar padrões não lineares e sutis tendem a apresentar melhor desempenho preditivo.

Algoritmos	F1_Score	Precisão	Sensibilidade (Recall)	Acurácia
MLP	92%	90%	95%	88,5%
Decision Tree Classifier	90%	90%	90%	85,5%
Extra Trees Classifier	83%	87%	80%	76,9%

Tabela II: Resultados a partir das métricas de avaliação.

A seguir, são apresentadas as matrizes de confusão de cada modelo, acompanhadas da análise de seu desempenho.

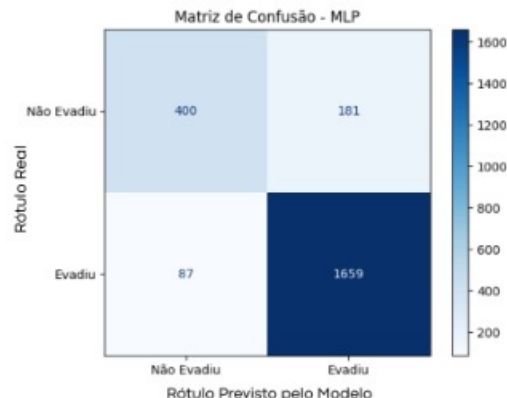


Figura 3: Matriz de confusão do modelo MLP.

Na Figura 3, do modelo MLP, foram encontrados os seguintes resultados:

- 1659 acertos — ao classificar que os estudantes evadiram (Verdadeiros Positivos);
- 181 erros — na previsão de evasão para estudantes que permaneceram no curso (Falsos Positivos);
- 400 acertos — na previsão de alunos que permaneceram (Verdadeiros Negativos);
- 87 erros — ao classificar que permaneceram estudantes que na verdade evadiram (Falsos Negativos).

A taxa baixa de falsos negativos é particularmente relevante do ponto de vista educacional, pois demonstra que o modelo é capaz de identificar a maioria dos alunos em risco de evasão.

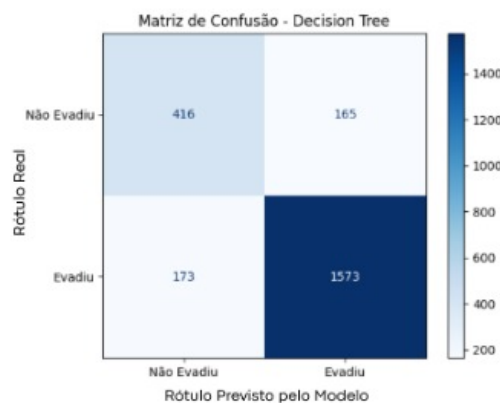


Figura 4: Matriz de confusão referente ao modelo Decision Tree.

- 1573 acertos — ao classificar que os estudantes evadiram (Verdadeiros Positivos);
- 165 erros — na previsão de evasão para estudantes que permaneceram no curso (Falsos Positivos);

- 416 acertos — na previsão de alunos que permaneceram (Verdadeiros Negativos);
- 173 erros — ao classificar que permaneceram estudantes que na verdade evadiram (Falsos Negativos).

O resultado da matriz de confusão tem uma boa taxa de acerto, mas ainda há espaço para melhorias na redução de falsos negativos, como pode ser observado na Figura 4, pois esses representam os estudantes em risco de evasão não identificados corretamente.

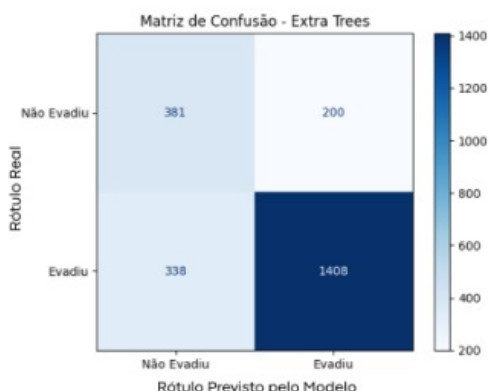


Figura 5: Matriz de confusão referente ao modelo Extra Trees.

- 1408 acertos — ao classificar que os estudantes evadiram (Verdadeiros Positivos);
- 200 erros — na previsão de evasão para estudantes que permaneceram no curso (Falsos Positivos);
- 381 acertos — ao classificar que os estudantes permaneceram (Verdadeiros Negativos);
- 338 erros — ao classificar que permaneceram estudantes que na verdade evadiram (Falsos Negativos).

O número elevado de falsos negativos exibidos na Figura 5 indica uma limitação importante do modelo em identificar estudantes em risco real de evasão. A partir da maior taxa de erro nessa classe, percebemos que o modelo tem menor sensibilidade (*Recall*) em comparação aos demais modelos.

VI. CONCLUSÃO

Este estudo utilizou técnicas de aprendizado de máquina para a predição da evasão acadêmica em cursos da área de Exatas do Instituto Federal de Pernambuco (IFPE). A análise foi realizada a partir de dados acadêmicos, demográficos e socioeconômicos, com o objetivo de identificar padrões e fatores que influenciam o risco de abandono escolar. Os resultados indicaram que o modelo de rede neural MLP apresentou maior capacidade preditiva, com taxa de acerto (acurácia) de 88,5%, precisão de 90%, recall de 95% e F1-Score de

92%. O modelo se destacou especialmente na identificação dos estudantes com maior probabilidade de evasão, devido à sua habilidade de lidar com relações complexas e não lineares entre as variáveis. O modelo Decision Tree Classifier, apesar de apresentar desempenho inferior, possibilitou uma melhor compreensão dos principais fatores associados à evasão, como o desempenho acadêmico e condições socioeconômicas. Neste estudo o foco foi descobrir se esses modelos poderiam ser úteis quando aplicados ao contexto da previsibilidade de evasão a partir dos dados do sistema acadêmico do IFPE. Os resultados obtidos indicam que há indícios fortes de que eles podem ser bastante efetivos para esse tipo de análise e que, portanto, podem contribuir significativamente para o desenvolvimento de políticas educacionais mais eficazes e para a redução dos índices de evasão no IFPE. Como trabalhos futuros, vemos não apenas a possibilidade de testar outros modelos, como a análise da influência de outros critérios, não considerados nesta análise, além de realizar uma análise com treinamentos dos modelos com dados de outras áreas, cursos, modalidades de ensino e instituições. Outra vertente é tentar encontrar um conjunto mínimo/ótimo de critérios que permita fazer uma previsão com percentuais de acerto ainda confiáveis.

REFERÊNCIAS

- [1] L. M. L. da Silva Garcia and R. S. Gomes, "Causas da evasão em cursos de ciências exatas: uma revisão da produção acadêmica," *Revista Educar Mais*, vol. 6, pp. 937–957, 2022, [Online]. Available: <https://periodicos.ifsul.edu.br/index.php/educarmais/article/view/2970>. Accessed: Feb. 21, 2025.
- [2] E. Souza and L. F. Freitas, "Um estudo sobre a evasão nos cursos de graduação dos institutos federais," *Revista Brasileira da Educação Profissional e Tecnológica*, vol. 1, no. 20, p. e10757, 2021.
- [3] A. Salata, "Razões da evasão: abandono escolar entre jovens no brasil," *Interseções: Revista de Estudos Interdisciplinares*, vol. 21, no. 1, 2019, [Online]. Available: <https://journals.openedition.org/intersecoes/5569>. Accessed: Jun. 11, 2025.
- [4] G. S. et al., "Siog2024-1-0a-004 predicting early mortality in older adults with stomach cancer: a machine learning (ml) approach incorporating comprehensive geriatric assessment (cga) and clinical data (cd)," *Journal of Geriatric Oncology*, vol. 15, no. 7, p. 101848, 2024.
- [5] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. García-Castelán, "Predictive analytics study to determine undergraduate students at risk of dropout," *Frontiers in Education*, vol. 8, 2023.
- [6] F. P. et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [7] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.
- [8] V. R. C. Martinho, C. Nunes, and C. R. Minussi, "A new method for prediction of school dropout risk group using neural network fuzzy artmap," in *Proc. Int. Conf. Artificial Intelligence (ICAI)*, 2013.
- [9] K. H. F. Kawase, "Aplicação de redes neurais rbf e mlp na análise de evasão discente do curso de sistemas de informação da ufrj," Master's thesis, Univ. Federal Rural do Rio de Janeiro, Instituto de Ciências Exatas, Seropédica, RJ, 2015.

- [10] S. Swaminathan and B. R. Tantri, “Confusion matrix–based performance evaluation metrics,” *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, 2024.