Desenvolvimento de um modelo de Inteligência Artificial com Scikit-Learn para Predição de Reprovação no Ensino Técnico Integrado do IFFar/FW

Amandha Ayres de Almeida[1], Árton Pereira Dorneles[1]

[1] Curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar) – Frederico Westphalen – RS – Brasil

amandha.2022002480@aluno.iffar.edu.br

arton.dorneles@iffarroupilha.edu.br

Abstract. *This paper presents a supervised machine learning model for predicting academic failure in technical education programs integrated with secondary education at the Instituto Federal Farroupilha – Frederico Westphalen Campus. The study is motivated by the role of academic failure as a factor associated with student dropout. Institutional academic data were prepared and analyzed using binary classification models implemented with the Scikit-learn library. The models were evaluated using stratified cross-validation and standard performance metrics. The results indicate that the proposed approach can support early identification of students at risk and assist academic management actions.*

Resumo. *Este trabalho apresenta um modelo de aprendizado de máquina supervisionado para a predição de reprovação escolar no ensino técnico integrado ao ensino médio do Instituto Federal Farroupilha – Campus Frederico Westphalen. O estudo é motivado pela relevância da reprovação como fator associado ao risco de evasão. Foram utilizados dados acadêmicos institucionais e modelos de classificação binária implementados com a biblioteca Scikit-learn. A avaliação foi realizada por meio de validação cruzada estratificada e métricas de desempenho. Os resultados demonstram o potencial da abordagem como ferramenta de apoio à gestão acadêmica.*

1. Introdução

A permanência e o êxito dos estudantes constituem indicadores fundamentais para a avaliação da qualidade da educação ofertada por instituições públicas, especialmente no âmbito da Rede Federal de Educação Profissional, Científica e Tecnológica. Nos Institutos Federais, esses indicadores estão diretamente associados à efetividade das políticas educacionais, à adequada aplicação dos recursos públicos e à promoção da formação integral dos alunos. Entretanto, a evasão escolar ainda representa um desafio significativo, sobretudo nos cursos técnicos integrados ao ensino médio, nos quais os estudantes enfrentam dificuldades de ordem socioeconômica, acadêmica e estrutural.

Entre os fatores que contribuem para a evasão, a reprovação escolar se destaca como um dos principais antecedentes, uma vez que o insucesso acadêmico tende a impactar diretamente a motivação, a autoestima e a permanência do estudante na instituição. De acordo com auditoria realizada pelo Tribunal de Contas da União (TCU), algumas unidades da Rede Federal apresentam taxas de evasão superiores a 24%, o que reforça a necessidade de ações institucionais voltadas não apenas ao combate direto da

evasão, mas também à identificação precoce de situações de risco acadêmico, como a reprovação (REZENDE; BRAGA; SOUSA, 2022).

Nesse contexto, ferramentas baseadas em Inteligência Artificial (IA) têm se consolidado como importantes aliadas no suporte à tomada de decisão em diferentes áreas, incluindo a educação. Por meio de técnicas de aprendizado de máquina, é possível analisar grandes volumes de dados históricos e identificar padrões associados ao desempenho acadêmico dos estudantes. Quando aplicadas ao contexto educacional, essas técnicas permitem a predição de reprovação escolar, possibilitando a atuação antecipada por parte da gestão acadêmica e das equipes pedagógicas, com potencial impacto na redução da evasão.

A literatura científica apresenta diversas abordagens que utilizam algoritmos de aprendizado supervisionado para a análise de dados educacionais. Trabalhos como os de Bezerra et al. (2016), Bitencourt, Silva e Xavier (2021) e Pimentel et al. (2023) demonstram a viabilidade do uso de técnicas como Regressão Logística, Árvores de Decisão e *Random Forest* para a identificação de estudantes em situação de risco acadêmico, a partir de variáveis como desempenho escolar, frequência e características socioeconômicas.

Diante desse cenário, este trabalho propõe o desenvolvimento de um modelo de aprendizado de máquina supervisionado para a predição de reprovação escolar no ensino técnico integrado do Instituto Federal Farroupilha – Campus Frederico Westphalen (IFFar/FW). Para a implementação do modelo e avaliação, foram utilizados algoritmos de classificação binária disponibilizados pela biblioteca Scikit-learn, aplicados a dados acadêmicos institucionais anonimizados. Ao antecipar situações de reprovação, espera-se que o modelo desenvolvido possa atuar como ferramenta de apoio à gestão acadêmica, contribuindo indiretamente para ações de permanência e êxito estudantil e podendo, futuramente, ser adaptado a outros campi da Rede Federal.

O restante deste trabalho está organizado como segue. A Seção 2 apresenta as principais definições, conceitos, tecnologias e trabalhos relacionados necessários para a melhor compreensão deste trabalho. A Seção 3 detalha a metodologia utilizada para o projeto do método proposto, bem como delimitação do escopo da pesquisa. A Seção 4 apresenta experimentos computacionais e análise de resultados. Por fim, a Seção 5 apresenta as principais conclusões desta pesquisa e possibilidades de trabalhos futuros.

2. Referencial teórico

2.1. Definições e Conceitos

2.1.1 Reprovação e Evasão nos Institutos Federais

No contexto da gestão acadêmica dos Institutos Federais, diversos indicadores são utilizados para monitorar o desempenho institucional, entre eles a taxa de reprovação, que permite identificar dificuldades no processo de ensino-aprendizagem e situações de risco na trajetória acadêmica dos estudantes. Esse indicador é amplamente utilizado na avaliação da efetividade das políticas educacionais e das condições de permanência estudantil.

A reprovação recorrente pode comprometer a progressão acadêmica do discente e contribuir para o aumento da evasão escolar, entendida como a interrupção do vínculo do estudante com o curso antes de sua conclusão. No âmbito da Rede Federal, estudos apontam como fatores associados a esse processo a vulnerabilidade socioeconômica, a defasagem na formação básica, a distância entre a residência e o campus e a necessidade de inserção precoce no mercado de trabalho, os quais tendem a intensificar os impactos negativos da reprovação.

Segundo auditoria realizada pelo Tribunal de Contas da União (TCU), a evasão nos cursos da Rede Federal alcançou índices de até 24% em determinados campi, evidenciando a necessidade de ações institucionais voltadas à identificação precoce de estudantes em situação de risco acadêmico. Nesse sentido, o monitoramento da reprovação torna-se estratégico, uma vez que possibilita intervenções preventivas que visem à permanência e à conclusão dos cursos (REZENDE; BRAGA; SOUSA, 2022).

2.1.2 Inteligência Artificial e Aprendizado de Máquina

A Inteligência Artificial (IA) é uma área da ciência da computação dedicada à criação de sistemas capazes de simular comportamentos inteligentes. As abordagens para defini-la variam entre aquelas que buscam imitar o pensamento humano, aquelas que focam em reproduzir o comportamento humano, e aquelas que se baseiam em ideais de racionalidade, agindo ou pensando de forma lógica e correta, dado um determinado conhecimento. Cada uma dessas visões têm contribuído para o desenvolvimento do campo, mesmo com diferentes métodos e objetivos (RUSSELL; NORVIG, 2021).

O aprendizado de máquina, por sua vez, é uma subárea da inteligência artificial que se concentra em construir sistemas capazes de aprender a partir da experiência humana. Em contraste com os sistemas tradicionais baseados em regras fixas, os algoritmos de aprendizado de máquina analisam grandes volumes de dados para reconhecer padrões e ajustar automaticamente seu desempenho. Isso permite que os modelos se tornem progressivamente mais precisos, de forma semelhante ao processo de aprendizado por tentativa e erro. Autores como Mitchell (1997) e Russell e Norvig (2021) destacam a importância desse tipo de abordagem, especialmente em contextos onde há dados históricos disponíveis. No cenário educacional, o uso de aprendizado de máquina tem potencial para prever o risco de reprovação dos estudantes a partir de seu histórico acadêmico, contribuindo para a identificação precoce de situações que podem levar à evasão e subsidiando ações estratégicas de intervenção.

O aprendizado de máquina pode ser dividido em três categorias principais: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, o modelo é treinado com dados rotulados — ou seja, cada entrada já possui uma saída conhecida — com o objetivo de encontrar uma função que consiga prever corretamente novos resultados. Essa abordagem é amplamente utilizada em tarefas como classificação e regressão (RUSSELL; NORVIG, 2021). Já o aprendizado não supervisionado utiliza dados não rotulados e busca identificar padrões ocultos, agrupamentos ou relações entre os dados. Um exemplo é o algoritmo K-means, que agrupa registros com características similares (CORCOVIA; ALVES, 2019). O aprendizado por reforço, por sua vez, é

baseado na interação de um agente com o ambiente, aprendendo por tentativa e erro a partir de recompensas e penalidades. Essa abordagem é comum em jogos ou simulações, como quando um sistema aprende estratégias vencedoras apenas com base nas vitórias e derrotas (RUSSELL; NORVIG, 2021). Especificamente neste projeto o foco será na utilização de técnicas supervisionadas, as quais serão detalhadas a seguir.

O aprendizado supervisionado é uma das formas mais comuns de ensinar um modelo de inteligência artificial a fazer previsões. Ele funciona com base em exemplos já conhecidos, onde cada entrada possui um rótulo associado. Entre as técnicas mais utilizadas estão a Regressão Logística, que calcula a chance de algo acontecer; a Árvore de Decisão, que organiza perguntas em sequência para tomar decisões; o Random Forest, que combina várias árvores para melhorar os resultados; e a Máquina de Vetores de Suporte (SVM), que separa os dados em grupos da forma mais eficiente possível. Os problemas de classificação podem ser mais simples, quando existem apenas duas categorias possíveis (classificação binária), ou mais complexos, quando é necessário escolher entre várias categorias (classificação multiclasse).

De modo geral, o desenvolvimento de um projeto com aprendizado supervisionado segue etapas que incluem a coleta e organização dos dados, a preparação e o pré-processamento, a escolha das informações mais relevantes, o treinamento do modelo, a avaliação do seu desempenho e, por fim, a possível implantação da solução. A preparação dos dados é uma etapa essencial, pois influencia diretamente na qualidade dos resultados obtidos. Esse processo envolve a correção de inconsistências, o tratamento de valores ausentes, a padronização de formatos e a transformação das variáveis, para que fiquem compatíveis com os algoritmos utilizados no treinamento. Para avaliar se o modelo está funcionando bem, são utilizadas métricas específicas. A acurácia mostra a proporção de previsões corretas no total. A precisão indica o quanto das previsões positivas estavam realmente certas. O recall mede quantos dos casos positivos reais foram corretamente identificados. Finalmente, a métrica F1-score representa um equilíbrio entre precisão e recall, oferecendo uma visão mais completa do desempenho do modelo.

Além das métricas de desempenho, a avaliação de modelos de aprendizado de máquina geralmente envolve técnicas que visam estimar sua capacidade de generalização. Entre essas técnicas, destaca-se a validação cruzada K-fold, amplamente utilizada em problemas de classificação. Nesse método, o conjunto de dados é particionado em K subconjuntos de tamanho semelhante, de forma que, a cada iteração, um desses subconjuntos é utilizado como conjunto de teste, enquanto os demais são empregados no treinamento do modelo. Esse processo é repetido K vezes, permitindo que todas as instâncias sejam utilizadas tanto para treinamento quanto para teste. A validação cruzada contribui para a obtenção de estimativas mais robustas do desempenho do modelo, reduzindo a dependência de uma única divisão dos dados e minimizando o risco de sobreajuste (CUNHA, 2019).

2.2. Tecnologias e Ferramentas

2.2.1 Python

O Python é uma linguagem de programação de alto nível, interpretada, com tipagem dinâmica e forte suporte à orientação a objetos. Foi desenvolvida com foco na legibilidade do código e na produtividade dos desenvolvedores (VAN ROSSUM; DRAKE, 2009). Atualmente, é amplamente adotada em áreas como ciência de dados, desenvolvimento web, automação e, especialmente, aprendizado de máquina. Entre suas principais vantagens estão a sintaxe simples, a ampla documentação e uma vasta gama de bibliotecas. Em ambientes educacionais e acadêmicos, destaca-se também por sua integração com plataformas como o Jupyter Notebook, que facilita a prototipagem e visualização de resultados.

2.2.2 Scikit-Learn

A Scikit-Learn é uma biblioteca open source desenvolvida em Python voltada ao aprendizado de máquina. Ela oferece ferramentas eficientes para mineração e análise de dados, sendo amplamente utilizada por pesquisadores e profissionais da área (PEDREGOSA, 2011). Além de disponibilizar implementações de algoritmos supervisionados como regressão logística, árvores de decisão, random forest e máquinas de vetor de suporte, também fornece algoritmos não supervisionados e recursos para validação cruzada, pipelines de modelagem, normalização e pré-processamento dos dados. A sua API simples e a compatibilidade com bibliotecas como NumPy, SciPy e Pandas tornam o Scikit-Learn ideal para construção de protótipos rápidos e experimentos controlados. Outro destaque é a documentação clara e objetiva, que favorece seu uso mesmo por iniciantes em *machine learning*.

2.3. Trabalhos relacionados

Três trabalhos acadêmicos foram analisados por apresentarem propostas semelhantes à deste projeto. Todos utilizam técnicas de mineração de dados e aprendizado de máquina para identificar estudantes em risco de reprovação, considerando-a como um fator associado ao risco de evasão, além de aplicarem metodologias semelhantes de coleta, preparação, modelagem e avaliação dos dados.

O primeiro trabalho, de Pimentel et al. (2023), apresenta uma abordagem para classificar alunos em risco de evasão no ensino superior por meio de atributos relacionados ao desempenho acadêmico e a fatores socioeconômicos. Nesse estudo, foram utilizados modelos como Regressão Logística, Árvores de Decisão e Random Forest, treinados a partir de históricos acadêmicos e variáveis socioeconômicas. Os autores observaram que fatores como baixo rendimento, reprovações anteriores e condições socioeconômicas desfavoráveis estão fortemente associados à evasão, e que os modelos alcançaram elevados níveis de acurácia, precisão e recall, demonstrando boa capacidade de distinguir estudantes em risco e não em risco.

Já em Bezerra et al. (2016), os autores analisaram a evasão no último ano do ensino fundamental em Pernambuco utilizando dados do Censo Escolar e aplicando técnicas como Árvores de Decisão e Regressão Logística. O estudo mostrou que variáveis como idade, turno, região geográfica e histórico escolar são relevantes para explicar a evasão, permitindo a construção de modelos capazes de identificar grupos de maior vulnerabilidade ao abandono escolar.

Finalmente, o estudo de Bitencourt, Silva e Xavier (2021) investigou a evasão universitária no IFMG utilizando algoritmos como Random Forest. Os autores desenvolveram um sistema de alerta baseado em inteligência artificial, no qual os modelos analisam o histórico acadêmico dos estudantes para indicar, de forma antecipada, aqueles com maior probabilidade de evasão, possibilitando a realização de intervenções institucionais direcionadas e preventivas.

3. Metodologia

Esta seção apresenta os procedimentos metodológicos adotados no desenvolvimento deste trabalho, descrevendo as etapas realizadas para a construção, treinamento e avaliação de modelos de aprendizado de máquina voltados à predição de reprovação escolar, bem como a caracterização do conjunto de dados utilizado na pesquisa.

3.1 Delimitação do Escopo dos Dados Educacionais

Devido à diversidade de informações presentes nos registros acadêmicos institucionais, tornou-se necessária a delimitação do escopo dos dados utilizados neste estudo. Foram analisados dados educacionais fornecidos pelo Instituto Federal Farroupilha (IFFar), Campus Frederico Westphalen, contemplando registros da trajetória de estudantes do 1º, 2º e 3º ano de um curso técnico não identificado do Ensino Médio Integrado.

O conjunto de dados utilizado é composto por registros anonimizados de estudantes, abrangendo três anos do curso. Para o 1º ano, foram utilizados 97 registros de estudantes, contendo 145 colunas de atributos acadêmicos. O 2º ano contou com 86 registros e 154 colunas, enquanto o 3º ano foi composto por 75 registros e 137 colunas. Esses atributos estão organizados em diferentes categorias, incluindo 7 colunas de dados cadastrais dos estudantes, 9 colunas por disciplina referentes às avaliações individuais e 12 colunas de notas globais e indicadores de desempenho, além de informações de frequência e situação final.

Os dados foram disponibilizados para esta pesquisa de forma anonimizada, garantindo a preservação da identidade dos estudantes e atendendo aos princípios éticos da pesquisa. O conjunto de dados incluiu informações relacionadas ao desempenho acadêmico consideradas relevantes para a análise da reprovação, como histórico de notas progressivo ao longo de cada ano, frequência escolar e situação de aprovação.

3.2 Procedimento de Preparação e Pré-processamento dos Dados

Os dados educacionais foram disponibilizados inicialmente em formato de planilhas eletrônicas contendo registros individuais dos estudantes e seus atributos acadêmicos. A etapa de preparação dos dados teve início com a seleção das características que seriam utilizadas no estudo, descartando-se atributos redundantes ou duplicados. Na sequência,

realizou-se a verificação da presença de valores nulos ou inconsistentes, bem como a adequação dos formatos das células de acordo com o tipo de dado de cada atributo. Atributos categóricos de natureza binária foram codificados numericamente, adotando-se o valor 1 para respostas afirmativas e 0 para negativas.

Também foi realizada a padronização dos nomes de cada atributo, especialmente aqueles pertencentes ao escopo de dados cadastrais e de notas globais, os quais foram convertidos para letras minúsculas e sem acentuação, com o objetivo de facilitar a manipulação dos dados no ambiente de programação. Os atributos contendo informações de disciplinas específicas mantiveram sua nomenclatura original, preservando informações como notas parciais, médias, frequência, faltas e situação final do estudante. Ao final desse processo foram gerados três arquivos no formato ".csv", correspondendo aos três anos do curso técnico integrado, e então usados como fonte de dados na implementação dos experimentos deste trabalho.

3.3 Treinamento dos Modelos com a Biblioteca Scikit-Learn

O treinamento dos modelos de aprendizado de máquina foi realizado utilizando a linguagem Python, com o apoio da biblioteca Scikit-learn, amplamente empregada no desenvolvimento de soluções de aprendizado supervisionado. Essa biblioteca foi selecionada por disponibilizar implementações consolidadas de algoritmos de classificação, além de recursos para pré-processamento, construção de pipelines e validação de modelos preditivos.

Inicialmente, os dados foram carregados a partir dos arquivos no formato .csv, sendo definidos os atributos preditores (*features*) e a variável alvo, correspondente à reprovação escolar. Os atributos foram organizados de forma progressiva, considerando diferentes momentos do ano letivo, permitindo avaliar a capacidade preditiva dos modelos à medida que novas informações acadêmicas se tornavam disponíveis.

Com o objetivo de representar marcos distintos do percurso acadêmico do estudante, foram definidos quatro momentos principais, descritos a seguir:

- **Momento 1:** corresponde ao período da primeira avaliação parcial, no qual estavam disponíveis apenas as notas iniciais das disciplinas e as informações cadastrais dos estudantes.
- **Momento 2:** representa o encerramento do primeiro semestre, incorporando as médias semestrais e informações adicionais relacionadas à frequência escolar.
- **Momento 3:** refere-se ao período da segunda avaliação parcial, incluindo novas notas parciais referentes ao segundo semestre letivo.
- **Momento 4:** corresponde ao encerramento do segundo semestre, antes da realização do exame final, contemplando médias finais, frequência acumulada e demais indicadores de desempenho acadêmico.

Essa estratégia possibilita analisar o desempenho dos modelos em diferentes estágios do ano letivo, permitindo identificar em qual momento a predição da reprovação escolar se torna mais precisa e, portanto, mais relevante para ações de acompanhamento e intervenção pedagógica.

Após a definição dos atributos, as variáveis preditoras foram separadas em numéricas e categóricas, possibilitando a aplicação de técnicas específicas de pré-processamento para cada tipo de dado. Para os atributos numéricos, foram aplicadas estratégias de imputação de valores ausentes pela média e normalização por meio da classe "StandardScaler". Para os atributos categóricos, utilizou-se imputação pelo valor mais frequente e codificação One-Hot por meio da classe "OneHotEncoder".

O processo de pré-processamento e treinamento foi estruturado por meio de *pipelines*, garantindo a aplicação consistente das transformações nos dados durante todas as etapas do experimento. Conforme ilustrado no trecho de código da Figura 1, foram definidos *pipelines* distintos para atributos numéricos e categóricos, os quais foram integrados por meio da classe "ColumnTransformer" e acoplados ao modelo de classificação em uma única estrutura. Essa abordagem assegura que todas as etapas de transformação sejam aplicadas de forma padronizada durante o treinamento e a validação dos modelos, evitando inconsistências entre os diferentes conjuntos de dados.

Para a etapa de classificação, a variável responsável pelo modelo (linha 101) é instanciada, de forma alternada, com as classes "RandomForestClassifier" e "MLPClassifier", ambas disponibilizadas pela biblioteca Scikit-learn, permitindo a comparação, respectivamente, de abordagens baseadas em árvores de decisão e redes neurais artificiais no contexto da predição de reprovação escolar.

```
83  pipeline_numerico = Pipeline(steps=[
84      ('imputer', SimpleImputer(strategy='mean')),
85      ('scaler', StandardScaler())
86  ])
87
88  pipeline_categorico = Pipeline(steps=[
89      ('imputer', SimpleImputer(strategy='most_frequent')),
90      ('onehot', OneHotEncoder(handle_unknown='ignore'))
91  ])
92
93  preprocessador = ColumnTransformer(
94      transformers=[
95          ('num', pipeline_numerico, colunas_numericas),
96          ('cat', pipeline_categorico, colunas_categoricas)
97      ]
98  )
99  clf = Pipeline([
100     ('preprocessador', preprocessador),
101     ('modelo', modelo)
102 ])
```

Figura 1 – Trecho de código com estrutura de pipelines de pré-processamento e treinamento.

3.4 Avaliação dos Modelos

A validação dos modelos foi realizada por meio da técnica de validação cruzada K-fold (StratifiedKFold), com cinco partições (K=5), assegurando a manutenção da proporção entre as classes em cada subconjunto. Para garantir a reprodutibilidade dos

experimentos, foi adotado um valor fixo de semente (random_state). A implementação deste procedimento é ilustrada no trecho de código apresentado na Figura 2.

```
105   skf = StratifiedKFold(n_splits=5, random_state=semente, shuffle=True)
106
107   print('\n*****************************************************')
108   print('Modelo', modelo.__class__.__name__)
109   print('Momento', momento)
110   print('*****************************************************')
111
112   f = 1
113   for index_treino, index_teste in skf.split(x, y):
114     print("\nFold: ", f)
115     f += 1
116
117     x_treino, x_teste = x.iloc[index_treino], x.iloc[index_teste]
118     y_treino, y_teste = y.iloc[index_treino], y.iloc[index_teste]
119
120     clf.fit(x_treino, y_treino)
121     y_prev = clf.predict(x_teste)
```

Figura 2 – Trecho de código da implementação da avaliação com validação cruzada.

Observando o código é possível notar que o conjunto de dados é dividido iterativamente em subconjuntos de treinamento e teste, constituindo um “fold”. Em cada "fold”, o modelo é treinado com os dados de treinamento e utilizado para gerar previsões sobre o subconjunto de dados de teste correspondente. As previsões obtidas em cada “fold”, servem de base para o cálculo das métricas de desempenho adotadas. Embora sejam coletadas acurácia, precisão, recall e F1-score, a métrica mais importante para este trabalho é o recall, pois com ela é possível quantificar de fato a parcela de estudantes em risco de reprovação que o modelo identifica corretamente.

4. Experimentos Computacionais e Resultados

Este capítulo apresenta os experimentos computacionais realizados e os resultados obtidos a partir da aplicação dos modelos de aprendizado de máquina propostos para a tarefa de predição de reprovação escolar. Inicialmente, são descritos o ambiente de execução e a configuração dos experimentos. Em seguida, são apresentados e analisados os resultados obtidos para cada ano do curso técnico integrado avaliado, considerando os diferentes momentos do percurso acadêmico dos estudantes.

4.1 Configuração dos Experimentos Computacionais

A implementação desenvolvida foi feita na linguagem de programação Python e os experimentos foram conduzidos na plataforma Google Colab, que oferece infraestrutura computacional adequada para execução de algoritmos de aprendizado de máquina. Foram utilizadas também bibliotecas amplamente consolidadas na área de ciência de dados e aprendizado de máquina, com destaque para o Scikit-learn, que apoiou a implementação dos algoritmos de classificação, pré-processamento, construção de pipelines e validação dos modelos.

Os experimentos foram realizados considerando diferentes anos do curso técnico integrado estudado (1º, 2º e 3º ano) e momentos do percurso acadêmico, conforme definidos na Seção 3.3. Foram conduzidos experimentos para cada combinação de ano, momento e modelo e, a seguir, os valores médios das métricas obtidas em cada fold da validação cruzada são reportados nas tabelas das seções seguintes. Além disso, foram avaliados dois modelos de aprendizado supervisionado para classificação binária: RandomForestClassifier (Árvores de Decisão) e MLPClassifier (Redes Neurais Artificiais).

4.2 Resultados para o 1º Ano do Curso Técnico Integrado

Nesta seção são apresentados os resultados obtidos para o 1º ano do curso técnico integrado, considerando os quatro momentos definidos ao longo do ano letivo. Esses resultados permitem analisar o comportamento dos modelos em um estágio inicial da formação acadêmica dos estudantes, no qual a quantidade de informações disponíveis aumenta progressivamente ao longo do período letivo.

O conjunto de dados referente ao 1º ano é composto por 97 registros de estudantes, dos quais aproximadamente 63% foram aprovados e 37% reprovados, caracterizando um cenário com moderado desbalanceamento entre as classes. Ao longo dos quatro momentos, o número de atributos utilizados pelos modelos cresce de 20 colunas no Momento 1 para 94 colunas no Momento 4, refletindo o acúmulo de informações acadêmicas ao longo do ano letivo. A Tabela 1 apresenta os resultados médios das métricas de avaliação para os modelos RandomForestClassifier (RFC) e MLPClassifier (MLPC), considerando os Momentos 1 a 4.

Tabela 1: Resultados médios no 1º ano

Momentos	Modelo	Acurácia	Precisão	Recall	F1-Score
Momento 1 (20 colunas)	RFC	78.37%	74.57%	66.79%	70.08%
	MLPC	69.05%	61.33%	52.86%	56.10%
Momento 2 (34 colunas)	RFC	79.26%	75.95%	69.29%	72.00%
	MLPC	73.26%	65.71%	61.43%	63.31%
Momento 3 (48 colunas)	RFC	80.32%	78.73%	69.64%	73.12%
	MLPC	73.26%	64.88%	64.29%	64.34%
Momento 4 (94 colunas)	RFC	83.42%	83.93%	72.14%	76.95%
	MLPC	81.42%	78.57%	69.64%	73.63%

De forma geral, observa-se que ambos os modelos apresentaram melhoria gradual no desempenho à medida que novas informações acadêmicas foram incorporadas ao conjunto de dados. Esse comportamento é coerente com o aumento do número de atributos disponíveis ao longo dos momentos, que passa de 20 colunas no Momento 1 para 94 colunas no Momento 4. No Momento 1, caracterizado pela disponibilidade limitada de informações, o modelo RFC apresentou desempenho superior ao MLPC em todas as métricas, destacando-se especialmente no recall e no F1-score. Considerando que, nesse ano, aproximadamente 37% dos 97 estudantes foram reprovados, o RFC conseguiu identificar uma parcela significativa desses casos mesmo

com um conjunto reduzido de variáveis. Esse resultado sugere maior robustez do RFC em cenários iniciais com dados ainda pouco consolidados.

Nos Momentos 2 e 3, com a inclusão das médias semestrais, informações de frequência e novas avaliações parciais, ambos os modelos apresentaram ganhos de desempenho, indicando que esses atributos contribuem positivamente para a tarefa de predição da reprovação escolar. Nesses momentos, o número de colunas aumenta para 34 e 48, respectivamente, ampliando o volume de informações utilizadas na classificação. Já no Momento 4, que representa o encerramento do segundo semestre antes do exame final, observa-se o melhor desempenho geral dos modelos, com destaque novamente para o RFC, que apresentou os maiores valores médios de acurácia e F1-score, utilizando um conjunto de 94 atributos, o que reforça o impacto da maior disponibilidade de dados na qualidade das predições.

4.3 Resultados para o 2º Ano do Curso Técnico Integrado

Os experimentos realizados com os dados do 2º ano do curso técnico integrado seguiram o mesmo procedimento experimental adotado para o 1º ano. O conjunto de dados deste ano é composto por 86 registros de estudantes, dos quais aproximadamente 74% foram aprovados e 26% reprovados. A Tabela 2 apresenta os resultados médios obtidos para os quatro momentos analisados.

Tabela 2: Resultados médios no 2º ano

Momentos	Modelo	Acurácia	Precisão	Recall	F1-Score
Momento 1 (19 colunas)	RFC	82.68%	82.00%	54.00%	60.58%
	MLPC	84.90%	72.00%	67.00%	68.51%
Momento 2 (34 colunas)	RFC	84.90%	85.33%	59.00%	64.45%
	MLPC	82.55%	75.83%	57.00%	59.81%
Momento 3 (49 colunas)	RFC	84.90%	76.67%	64.00%	65.78%
	MLPC	79.22%	65.24%	59.00%	59.86%
Momento 4 (95 colunas)	RFC	88.37%	82.00%	68.00%	72.59%
	MLPC	78.04%	59.22%	68.00%	61.86%

Os resultados indicam que, de modo geral, ambos os modelos apresentaram desempenho superior em comparação ao 1º ano, especialmente em termos de acurácia. Esse comportamento pode estar associado tanto à menor proporção de reprovações quanto ao maior histórico acadêmico disponível no 2º ano, o que contribui para uma melhor capacidade de discriminação entre as classes. Além disso, o aumento progressivo do número de atributos, que vai de 19 colunas no Momento 1 até 95 colunas no Momento 4, fornece aos modelos informações mais completas sobre o desempenho dos estudantes.

Observa-se também que, embora a acurácia dos modelos seja elevada, as métricas de recall e F1-score apresentam variações mais expressivas entre os momentos, indicando desafios adicionais na identificação correta dos estudantes reprovados. Esse efeito é esperado em um cenário no qual apenas cerca de um quarto dos estudantes

pertence à classe de reprovação, o que torna a detecção desses casos mais sensível às variações nos dados. Ainda assim, o modelo RFC manteve desempenho consistente ao longo dos quatro momentos, enquanto o MLPC apresentou maior variação entre as métricas avaliadas.

4.4 Resultados para o 3º Ano do Curso Técnico Integrado

De forma análoga, os experimentos foram conduzidos com os dados do 3º ano do curso técnico integrado, permitindo analisar o comportamento dos modelos em um estágio mais avançado da trajetória acadêmica dos estudantes. O conjunto de dados desse ano é composto por 75 registros, dos quais aproximadamente 81% foram aprovados e apenas 19% reprovados, caracterizando um cenário fortemente desbalanceado. A Tabela 3 apresenta os resultados médios obtidos para os quatro momentos avaliados.

Tabela 3: Resultados médios no 3º ano

Momentos	Modelo	Acurácia	Precisão	Recall	F1-Score
Momento 1 (19 colunas)	RFC	82.67%	30.00%	13.33%	18.00%
	MLPC	78.67%	25.00%	13.33%	15.71%
Momento 2 (32 colunas)	RFC	85.33%	45.00%	33.33%	35.14%
	MLPC	80.00%	46.67%	33.33%	38.67%
Momento 3 (44 colunas)	RFC	86.67%	45.00%	40.00%	41.14%
	MLPC	77.33%	36.67%	26.67%	30.67%
Momento 4 (86 colunas)	RFC	89.33%	75.00%	50.00%	56.48%
	MLPC	74.67%	35.00%	20.00%	23.71%

Embora os valores de acurácia observados sejam elevados, nota-se que as métricas de precisão, recall e F1-score, especialmente nos Momentos 1 e 2, apresentam valores significativamente inferiores quando comparadas aos anos anteriores. Esse comportamento está diretamente relacionado à baixa proporção de estudantes reprovados no 3º ano, o que dificulta a aprendizagem dos padrões associados à classe minoritária, sobretudo nos estágios iniciais, quando apenas 19 colunas (Momento 1) e 32 colunas (Momento 2) estão disponíveis.

No Momento 4, observa-se uma melhora expressiva no desempenho do modelo RFC, especialmente nas métricas de precisão e F1-score, indicando que a maior disponibilidade de informações acadêmicas ao final do ano letivo, representada pelo uso de 86 atributos, contribui para uma predição mais assertiva da reprovação escolar nesse contexto.

4.5 Análise Geral dos Resultados com Foco no Recall

Uma vez que o objetivo central deste trabalho é a identificação antecipada de estudantes em risco de reprovação, o recall constitui a métrica mais relevante para a interpretação do desempenho dos modelos propostos. Diferentemente da acurácia global, que pode mascarar o desempenho em cenários com desbalanceamento entre as classes, o recall

expressa diretamente a capacidade do modelo em identificar corretamente os estudantes que de fato irão reprovar. Nesse sentido, é importante uma análise separada focada apenas no recall e que será apresentada a seguir.

Observando os resultados obtidos para o 1º e 2º anos, os modelos alcançaram valores de recall próximos de dois terços dos casos de reprovação já nos momentos iniciais do ano letivo, indicando que aproximadamente 6 a 7 em cada 10 estudantes reprovados podem ser identificados de forma antecipada. Esse resultado é particularmente relevante do ponto de vista da gestão acadêmica, pois indica que o uso do método proposto permitiria a adoção de ações de acompanhamento e intervenção pedagógica ainda em estágios iniciais do percurso acadêmico dos estudantes.

Também pode ser observado que, à medida que novas informações acadêmicas são disponibilizadas ao longo do ano letivo, existe uma tendência de manutenção ou melhoria do recall, especialmente nos Momentos 2 e 3. No Momento 4, os modelos atingem seus melhores valores de recall, evidenciando que a quantidade crescente de informações (colunas) usadas no treinamento dos modelos contribui para a melhoria da capacidade preditiva. Embora o Momento 4 represente um estágio mais tardio para intervenções preventivas, os resultados confirmam a consistência da abordagem proposta e validam o potencial dos modelos para apoiar decisões acadêmicas em diferentes momentos do ano letivo.

Em contraste, os resultados da métrica recall obtidos para os dados do 3º ano devem ser interpretados com cuidado visto que a baixa proporção de estudantes reprovados nesse ano caracteriza um conjunto desbalanceado nos dados, impactando negativamente a capacidade dos modelos em identificar reprovações, especialmente nos momentos iniciais, nos quais a quantidade de informações disponíveis é menor. Desta forma, os resultados indicam que, no 3º ano, os modelos não são adequados para fins de predição antecipada.

Em resumo, os resultados obtidos pela métrica recall demonstram que a abordagem proposta apresenta maior efetividade nos anos iniciais do curso técnico integrado, alinhando-se ao objetivo deste trabalho de apoiar a gestão acadêmica na promoção da permanência e do êxito estudantil com uma detecção precoce de estudantes em risco de reprovação.

5. Considerações Finais

Este trabalho teve como objetivo desenvolver e avaliar um modelo de Inteligência Artificial, utilizando técnicas de aprendizado de máquina com a biblioteca Scikit-learn, para a predição de reprovação escolar no ensino técnico integrado do Instituto Federal Farroupilha – Campus Frederico Westphalen. A proposta buscou investigar a viabilidade do uso de dados acadêmicos históricos como subsídio para a identificação precoce de estudantes em risco de reprovação, contribuindo para ações de acompanhamento e intervenção pedagógica.

Para atingir esse objetivo, foram utilizados dados educacionais anonimizados referentes à trajetória acadêmica completa de estudantes do 1º, 2º e 3º anos de um curso técnico integrado não identificado. Os dados passaram por um processo de preparação e

pré-processamento, contemplando seleção de atributos relevantes, tratamento de valores ausentes, padronização e organização em quatro momentos distintos do ano letivo, o que permitiu analisar a evolução do desempenho preditivo dos modelos à medida que novas informações acadêmicas se tornavam disponíveis.

Os experimentos computacionais foram conduzidos utilizando os classificadores RandomForestClassifier e MLPClassifier, avaliados por meio da técnica de validação cruzada com cinco partições e métricas de desempenho amplamente consolidadas na literatura, como acurácia, precisão, recall e F1-score. Considerando o objetivo central do trabalho, a análise dos resultados evidenciou a relevância do recall como métrica prioritária, uma vez que ela expressa a capacidade dos modelos em identificar corretamente os estudantes que efetivamente irão reprovar.

Os resultados obtidos demonstraram que, especialmente nos dados referentes ao 1º e ao 2º anos do curso técnico integrado, os modelos foram capazes de identificar uma parcela significativa dos estudantes em risco de reprovação já nos momentos iniciais do ano letivo. Em particular, observou-se que o classificador RandomForestClassifier alcançou valores de recall próximos a 70% nesses cenários, indicando potencial para subsidiar intervenções pedagógicas antecipadas, quando ainda há maior potencial para acompanhamento e apoio institucional aos alunos em risco.

De modo geral, o classificador RandomForestClassifier apresentou desempenho superior ao MLPClassifier na maioria dos cenários avaliados, destacando-se principalmente nos momentos iniciais do ano letivo e em contextos com maior complexidade dos dados. Observou-se também que o Momento 4, correspondente ao encerramento do segundo semestre antes do exame final, concentrou os melhores resultados médios para todos os anos analisados, indicando que a maior disponibilidade de informações acadêmicas contribui significativamente para a precisão das predições.

Por outro lado, os resultados obtidos para o 3º ano do curso técnico integrado evidenciaram limitações importantes relacionadas ao significativo desbalanceamento entre as classes, o que impactou negativamente o desempenho dos modelos, especialmente em termos de recall nos momentos iniciais. Nesse contexto, embora a acurácia global tenha se mantido elevada, nesse estágio em particular do curso, os modelos não apresentam utilidade prática para fins de intervenção precoce.

Os resultados reforçam, portanto, o potencial do uso de técnicas de aprendizado de máquina como ferramenta de apoio à gestão acadêmica, possibilitando a identificação de estudantes em risco de reprovação de forma antecipada, sobretudo nos anos iniciais. A aplicação prática de modelos como os desenvolvidos neste trabalho pode auxiliar instituições de ensino na tomada de decisões estratégicas, promovendo ações de acompanhamento pedagógico mais direcionadas e contribuindo para a permanência e êxito dos estudantes.

Como limitações do estudo, destacam-se a dependência da qualidade e da distribuição dos dados disponíveis, bem como possíveis efeitos de desbalanceamento entre as classes, especialmente nos anos finais do curso. Além disso, o trabalho foi

restrito a dois algoritmos de classificação, o que abre espaço para comparações futuras com outras abordagens.

Como trabalhos futuros, sugere-se a ampliação do conjunto de algoritmos de aprendizado de máquina avaliados, possibilitando a comparação com outras abordagens de classificação e a investigação de modelos com diferentes características e capacidades de generalização. Recomenda-se também a expansão e diversificação da base de dados utilizada, incorporando um maior volume de registros e novos atributos acadêmicos, o que pode contribuir para o aprimoramento do desempenho preditivo e para a obtenção de resultados mais robustos e representativos. Além disso, destaca-se a possibilidade de integração do modelo desenvolvido a sistemas acadêmicos institucionais, permitindo o acompanhamento contínuo dos estudantes ao longo do ano letivo e o apoio à tomada de decisão por parte da gestão pedagógica.

Referências

BEZERRA, C. et al. Evasão escolar: aplicando mineração de dados para identificar variáveis relevantes. In: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE), 2016, Recife. Anais [...]. Recife: Sociedade Brasileira de Computação, 2016. Disponível em: https://doi.org/10.5753/cbie.sbie.2016.1096 . Acesso em: 16 maio 2025.

BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 30, n. 116, p. 669–694, 2021. Disponível em: https://doi.org/10.1590/S0104-403620220003002854 . Acesso em: 16 maio 2025.

BRASIL. Lei nº 11.892, de 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica, cria os Institutos Federais de Educação, Ciência e Tecnologia, e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, n. 251, p. 1, 30 dez. 2008.

CORCOVIA, L. O.; ALVES, R. S. Aprendizagem de máquina e mineração de dados: avaliação de métodos de aprendizagem. Revista Interface Tecnológica, Taquaritinga, v. 16, n. 1, p. 90–101, 2019. Disponível em: https://revista.fatectq.edu.br/interfacetecnologica/article/view/562 . Acesso em: 24 maio 2025.

CUNHA, J. P. Z. Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. 2019. Dissertação (Mestrado em Ciências) – Universidade de São Paulo, São Paulo, 2019.

GONÇALVES, A. C. et al. A evasão no ensino superior público brasileiro: desafios e alternativas. Revista Brasileira de Educação, Rio de Janeiro, v. 24, p. 1–20, 2019. Disponível em: https://www.scielo.br/j/rbedu/a/JN9RwrFTrdVpbdS6Pgfb3Jg/ . Acesso em: 16 maio 2025.

LOBO, F. W. B. H.; SILVA, L. C.; FERREIRA, A. C. Fatores de risco associados à evasão escolar no Brasil. Educação e Sociedade, Campinas, v. 28, n. 100, p. 1015–1030, 2007.

MITCHELL, T. M. Machine learning. New York: McGraw-Hill, 1997. Disponível em: https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf . Acesso em: 17 maio 2025.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, [S. l.], v. 12, p. 2825–2830, 2011. Disponível em: https://jmlr.org/papers/v12/pedregosa11a.html . Acesso em: 16 maio 2025.

PIMENTEL, M. S. et al. Classificação de alunos em risco de evasão escolar utilizando modelos de machine learning. Apoena – Revista Eletrônica, Salvador, v. 7, p. 81–97, 2023. Disponível em: https://transformauj.com.br/apoena-revista-eletronica . Acesso em: 16 maio 2025.

REZENDE, S. E. de F.; BRAGA, R. B.; SOUSA, M. de M. Fatores de evasão escolar nos cursos técnicos integrados ao ensino médio: protocolo de revisão sistemática da literatura. Revista Temas em Educação, João Pessoa, v. 31, n. 2, p. 89–104, maio/ago. 2022. Disponível em: https://doi.org/10.22478/ufpb.2359-7003.2022v31n2.62371 . Acesso em: 17 maio 2025.

RUSSELL, S.; NORVIG, P. Inteligência artificial. 3. ed. Porto Alegre: AMGH, 2021.

SARKAR, D.; MURPHY, A.; BALCH, T. Practical machine learning with Python. Birmingham: Packt Publishing, 2018.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE. Análise e modelagem para prevenção de evasão escolar. 2024. Monografia (Graduação) – Universidade Federal do Rio Grande do Norte, Natal, 2024. Disponível em: https://repositorio.ufrn.br/ . Acesso em: 17 maio 2025.

VAN ROSSUM, G.; DRAKE, F. L. The Python language reference manual. Scotts Valley: CreateSpace, 2009.