UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

PEDRO MIAN PARRA

PREDIÇÃO DE EVASÃO ESTUDANTIL USANDO APRENDIZADO DE MÁQUINA

APUCARANA

2024

PEDRO MIAN PARRA

PREDIÇÃO DE EVASÃO ESTUDANTIL USANDO APRENDIZADO DE MÁQUINA

Student Dropout Prediction Using Machine Learning

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Rafael Gomes Mantovani

APUCARANA

2024

PEDRO MIAN PARRA

PREDIÇÃO DE EVASÃO ESTUDANTIL USANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 19/12/2024

Prof. Dr. Edgar De Souza Vismara
Doutorado
Universidade Tecnológica Federal do Paraná

Prof. Dr. Luiz Fernando Carvalho
Doutorado
Universidade Tecnológica Federal do Paraná

Prof. Dr. Rafael Gomes Mantovani
Doutorado
Universidade Tecnológica Federal do Paraná

APUCARANA

2024

Dedico este trabalho aos meus pais, em agradecimento pelo imenso esforço e dedicação que sempre depositaram na educação de seus três filhos.

AGRADECIMENTOS

Me graduar em Engenharia de Computação foi um sonho meu acolhido por diversas pessoas ao meu redor, as quais desejavam, acima de tudo, me ver feliz. Ao concluir este trabalho e olhar para trás, sou grato por reconhecer a importância que cada uma teve para que eu chegasse até aqui.

Agradeço primeiramente a Deus, que me concedeu força, consolo e perseverança, permitindo que eu superasse cada desafio e avançasse em direção ao meu objetivo.

Aos meus pais, Marcelo e Tânia, que, inúmeras vezes, abriram mão de seus próprios desejos para garantir que eu tivesse a melhor educação possível. Aos meus irmãos, Thiago e Lucas, que abriram caminhos e me mostraram como ir mais longe e superar as dificuldades. À toda minha família, em especial aos meus padrinhos Alexandre e Luciana, e a Maria Carolina.

Aos meus queridos amigos de Penápolis, com quem compartilhei o sonho de ingressar na faculdade, e juntos enfrentamos as dificuldades do vestibular, sempre nos apoiando e nos motivando a seguir em frente.

Aos amigos que fiz durante a graduação, e que se tornaram uma fonte constante de apoio, motivação e conforto durante essa jornada desafiadora. Em muitos momentos, vocês me trouxeram alento, paz e a força necessária para continuar.

Ao Prof. Dr. Rafael Gomes Mantovani, cuja orientação, sabedoria e inspiração foram essenciais para minha jornada acadêmica e para o meu crescimento profissional. À toda coordenação do curso de Engenharia de Computação da UTFPR Campus Apucarana, em especial ao coordenador, Prof. Dr. Luiz Fernando Carvalho, cuja liderança e dedicação à excelência acadêmica contribuíram para o meu aprendizado durante essa trajetória.

A cada um de vocês que, de forma única e especial, contribuíram para que eu pudesse concluir essa etapa tão importante da minha vida, minha eterna gratidão.

“Talvez não tenha conseguido fazer o melhor,
mas lutei para que o melhor fosse feito. Não
sou o que deveria ser, mas Graças a Deus, não
sou o que era antes”. (Marthin Luther King)

RESUMO

Nos últimos anos, a evasão no Ensino Superior tem se intensificado, gerando impactos negativos tanto para os estudantes quanto para as instituições, especialmente no setor público, que sofre perdas financeiras significativas. Diversos fatores contribuem para esse fenômeno, como dificuldades de aprendizado, estrutura inadequada dos cursos e a falta de recursos financeiros. O presente estudo propõe utilizar algoritmos de Aprendizado de Máquina (AM) para prever e identificar padrões comportamentais de estudantes com risco de evasão. Para isso, foram realizados experimentos com dados extraídos do sistema acadêmico da Universidade Tecnológica Federal do Paraná (UTFPR), campus Dois Vizinhos. Os dados foram devidamente anonimizados e pré-processados, para que informações sensíveis não fossem acessadas pelos modelos preditivos. Foram realizados experimentos com diferentes tarefas preditivas explorando toda a completude dos registros extraídos do sistema da universidade. Explorou-se também quatro algoritmos "interpretáveis", que retornam regras das estruturas que permitem os gestores entenderem as predições e tomarem decisões de gestão. Os resultados experimentais mostraram que o melhor algoritmo avaliado foi o *Random Forest*, com melhor valor médio de *F1-Score* em todos os experimentos, e atingindo o valor máximo de $0{,}99$ considerando o dataset com apenas os últimos registros dos alunos no sistema acadêmico, na classificação binária entre alunos “Regulares vs Desistentes” com limiares de atributos constantes de $0{,}05$ e correlacionados de $0{,}8$. Embora os resultados sejam promissores, eles ainda carecem da inclusão de variáveis socioeconômicas, o que limita a capacidade do sistema em compreender plenamente as causas da evasão.

Palavras-chave: aprendizado de máquina; evasão estudantil; predição de evasão.

ABSTRACT

In recent years, dropout rates in Higher Education have intensified, causing negative impacts on students and institutions, particularly in the public sector, which suffers significant financial losses. Various factors contribute to this phenomenon, such as learning difficulties, inadequate course structures, and a lack of financial resources. This study proposes using Machine Learning (ML) algorithms to predict and identify behavioral patterns of students at risk of dropping out. For this purpose, experiments were conducted using data extracted from the academic system of the Federal Technological University of Paraná (UTFPR), Dois Vizinhos campus. The data were properly anonymized and pre-processed to ensure that sensitive information could not be accessed by the predictive models. Experiments were conducted with different predictive tasks, exploring the completeness of the records extracted from the university's system. Additionally, four "interpretable"algorithms were explored, which provide rules or structures that allow managers to understand predictions and make informed management decisions. The experimental results showed that the best-performing algorithm was Random Forest, achieving the highest average F1-Score across all experiments, with a maximum value of 0.99. This result was obtained using the dataset containing only the latest student records from the academic system, considering a binary classification between "Regular vs. Dropout" students, with attribute thresholds of 0.05 for constants and 0.8 for correlations. Although the results are promising, they still lack the inclusion of socioeconomic variables, which limits the system's ability to understand the causes of dropout fully.

Keywords: machine learning; student dropout; dropout prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

BAC	Acurácia Balanceada, do inglês *Balanced Accuracy*
CD	Diferença Crítica, do inglês *Critical Difference*
DT	Árvore de Decisão, do inglês *Decision Tree*
EDM	Mineração de Dados Educacionais, do inglês *Educational Data Mining*
KNN	K-vizinhos mais próximos, do inglês *K-Nearest Neighbors*
MLP	*Multi-layer Perceptron*
NB	*Naive Bayes*
RF	Floresta Aleatória, do inglês *Random Forest*
SVM	Máquinas de Vetores de Suporte, do inglês *Support Vector Machine*
VM	Máquina Virtual, do inglês *Virtual Machine*
ABRUEM	Associação Brasileira dos Reitores das Universidades Estaduais e Municipais
AM	Aprendizado de Máquina
ANDIFES	Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior
AVA	Ambiente Virtual de Aprendizagem
CEP	Comitê de Ética em Pesquisa
EaD	Educação a Distância
ENEM	Exame Nacional do Ensino Médio
FN	Falso-Negativos
FP	Falso-Positivos
IA	Inteligência Artificial
IDEs	Interfaces de Desenvolvimento Integrado
IE	Instituição de Ensino Superior

IEs	Instituições de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
RN	Redes Neurais
RNA	Redes Neurais Artificiais
SESU	Secretaria de Educação Superior
UFJF	Universidade Federal de Juiz de Fora
UFSCar	Universidade Federal de São Carlos
UTFPR	Universidade Tecnológica Federal do Paraná
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

SUMÁRIO

1 INTRODUÇÃO

A evasão estudantil tem sido um problema persistente por muitos anos em diversas instituições de ensino e cursos universitários, acarretando em uma série de desafios e interferindo na gestão universitária (Ribeiro, 2005). Ela é definida por diversos autores como a interrupção abrupta da jornada educacional de um aluno, antes da conclusão do seu curso (Baggi; Lopes, 2011). Diante desse desafio macro, surge a necessidade de explorar de forma adequada os dados relacionados, visando extrair informações valiosas para auxiliar na mitigação deste problema. Paralelo a este contexto, o Aprendizado de Máquina (AM) emerge como uma ferramenta poderosa e revolucionária, transformando a maneira que lidamos com uma variedade de situações em nosso cotidiano (Alpaydin, 2021). A análise de grandes volumes de dados e a identificação de padrões complexos pelo AM são tarefas que possibilitam a otimização de processos e a tomada de decisões mais inteligentes em diversos setores.

Devido a esta capacidade de identificar padrões, o AM tem sido fortemente utilizado para analisar e prever o comportamento dos estudantes, como em Vossen *et al.* (2023) e Solis *et al.* (2018). Este movimento deu origem a um campo de pesquisa denominado Mineração de Dados Educacionais, do inglês *Educational Data Mining* (*EDM*), onde técnicas de mineração de dados são utilizadas para descobrir padrões e tendências em dados educacionais (Kabathova; Drlik, 2021). No âmbito internacional, existem várias pesquisas expressivas que se baseiam no uso do *EDM* (Manhães; Cruz; Zimbrão, 2014). Já aqui no Brasil, diversos autores tem utilizado dados educacionais para auxiliar na melhoria do ensino superior e na amenização da evasão estudantil. Alguns exemplos são os trabalhos de Silva (2022), que através da base de dados disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) buscou classificar a situação dos estudantes; e Teodoro e Kappel (2020) e Paz e Cazella (2017) que buscaram identificar as principais características dos estudantes que interrompiam a sua graduação, e quais eram as informações cruciais para chegar a essa conclusão.

Assim, a hipótese de pesquisa que rege este trabalho é de que, por meio do uso de algoritmos de AM, é possível identificar padrões distintivos nos dados acadêmicos e sociodemográficos que predizem a evasão estudantil. O estudo proposto busca analisar e explorar os dados históricos e contemporâneos dos alunos, presentes no Sistema Acadêmico da Universidade Tecnológica Federal do Paraná (UTFPR) no Campus de Dois Vizinhos. Além disso, supõe que os modelos obtidos serão capazes de rotular com precisão a situação de um aluno, classificando-o como evadido ou não evadido, podendo oferecer suporte a gestores, coordenadores e reitores na criação de políticas estudantis que minimizem esse problema.

1.1 Objetivos

1.1.1 Objetivo gerais

O objetivo geral deste trabalho é aplicar e avaliar o uso de algoritmos de AM para identificar estudantes evadidos na UTFPR, focando no Campus de Dois Vizinhos.

1.1.2 Objetivos específicos

Especificamente, pretende-se:

- Realizar uma análise exploratória nos dados obtidos do sistema acadêmico, com o objetivo de observar as principais características dos alunos evasores, bem como os principais grupos em que ocorre este comportamento, além das correlações e constâncias entre os atributos presentes na base de dados;
- Aplicar algoritmos de classificação interpretáveis, como *DT* e *RF*, para identificar o modelo mais adequado na classificação dos perfis de alunos evasores, considerando métricas de desempenho e avaliação;
- Investigar a importância de cada variável nos modelos gerados, destacando os principais padrões comuns entre os alunos evasores;
- Desenvolver um *pipeline* automatizado para o treinamento de modelos voltados à identificação de alunos evasores, explorando diferentes técnicas de pré-processamento de dados e algoritmos de AM.

1.2 Justificativa

Ao analisar a evasão estudantil, entende-se ela como um problema social em razão das diversas consequências e ramificações que ela acarreta. Entre essas, destaca-se a perda de recursos financeiros para instituições públicas, que acabam não obtendo o retorno dos investimentos públicos realizados, bem como a diminuição da receita no setor privado (Filho *et al.*, 2007). A longo prazo, as implicações sociais podem ser significativas, tanto para o indivíduo como para a sociedade. Evangelista (2017) pontua que o aluno que não recebe o diploma, pode não conseguir a sua inserção no mercado de trabalho e, consequentemente, pode não contribuir para o desenvolvimento regional.

É cada vez mais claro que a aplicação de métodos de análise de dados em ambientes educacionais proporciona oportunidades valiosas para educadores e pesquisadores. Através dessas técnicas e algoritmos, é possível descobrir quais comportamentos e decisões contri-

buem para o sucesso dos alunos, identificar aqueles que correm o risco de desistir ou apresentar desempenho abaixo do esperado, personalizar o conteúdo e a instrução para atender às necessidades específicas de cada aluno, e aprimorar a utilização dos recursos educacionais de forma mais eficiente (Rodrigues, 2018).

O uso do AM tem potencial para auxiliar na mitigação do problema da evasão estudantil, podendo ser aplicado aos diversos dados existentes dos alunos, e permitindo que as instituições de ensino intervenham precocemente, oferecendo suporte personalizado e recursos adicionais aos alunos que necessitarem. Os algoritmos de AM vêm sendo utilizados em grande escala nas indústrias, na área da saúde e em diversos outros campos (Ludermir, 2021). Sua capacidade de analisar grandes volumes de dados têm gerado benefícios gigantescos para quem os utiliza. No campo da educação não é diferente: o uso da Inteligência Artificial (IA) pode trazer milhares de benefícios, como mostra Cardoso *et al.* (2023), podendo apoiar o progresso de uma educação mais eficiente, atrativa e apta a suprir as exigências de uma sociedade cada vez mais conectada.

Este trabalho consistirá na modelagem e implementação de modelos de AM para prever a evasão estudantil, com um *pipeline* que inclui desde a coleta e preparação dos dados relevantes, seleção e treinamento dos algoritmos de AM mais adequados, e a avaliação de desempenho e eficácia dos modelos desenvolvidos. O presente trabalho focará no uso de algoritmos de classificação para identificar padrões nos dados e desenvolver modelos preditivos.

1.3 Estrutura do Texto

Este trabalho está organizado em mais cinco capítulos. No Capítulo 2, é apresentado o referencial teórico do trabalho, explorando os conceitos que envolvem a evasão e o AM. No Capítulo 3, encontra-se a análise dos trabalhos relacionados, explicitando os estudos que utilizam o AM para prever a evasão estudantil. No Capítulo 4, é descrita toda a metodologia experimental deste trabalho, desde a aquisição dos dados, pré-processamento, até os algoritmos usados para realizar as predições. Os resultados experimentais são reportados no Capítulo 5, analisando o desempenho preditivo dos algoritmos de AM e diferentes tarefas preditivas variando os dados descritivos. E, por fim, são apresentadas as conclusões do trabalho, limitações, e possibilidades de trabalhos futuros no Capítulo 6.

2 REFERENCIAL TEÓRICO

O presente capítulo introduz a fundamentação teórica necessária para a compreensão deste trabalho.

2.1 Evasão

A evasão estudantil é um fenômeno social complexo e abrangente, gerando diversos impactos na gestão universitária, além de ser influenciada por diversos fatores (Fialho, 2014). Logo, é fundamental entender o seu conceito antes de analisá-la. É preciso garantir que os modelos de AM serão implementados e treinados com informações relevantes e significativas. Com essa concepção de evasão estudantil, será possível alcançar a interpretação dos resultados que serão obtidos.

2.1.1 Definição

Em 1995 foi criada a Comissão Especial sobre Estudos de Evasão, ligada a Secretaria de Educação Superior (SESU) do Ministério da Educação (MEC), com o objetivo de estudar a fundo o tema no ensino superior. Em 1996, o relatório do MINISTÉRIO DA EDUCAÇÃO (1996) elaborado pela comissão foi apresentado à Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES), à Associação Brasileira dos Reitores das Universidades Estaduais e Municipais (ABRUEM) e à SESU, onde o termo evasão foi definido como a saída do aluno do seu curso, sem concluí-lo, podendo ter três diferentes níveis:

- **Evasão de curso:** sendo o desligamento do estudante do curso superior por abandono, desistência, transferência ou reopção, exclusão por norma institucional;
- **Evasão da instituição:** estudante que se desliga da instituição em que está matriculado;
- **Evasão do sistema:** estudante que abandona de forma definitiva ou temporária o ensino superior.

Referente a essas definições, os autores Silva, Cabral e Pacheco (2016) pontuam que não é possível que o aluno evada de uma dimensão macro sem evadir consequentemente de dimensões menores. Em outras palavras, quando o aluno evade do sistema de ensino, consequentemente é enquadrado como evadido da instituição e evadido do curso. Já quando o aluno opta por transferência interna, pode ser caracterizado apenas como evadido do curso, por continuar na mesma instituição de ensino e permanecer no sistema educacional. No estudo, é analisado também o conceito de temporaneidade da evasão, podendo ser definitiva ou temporária, pois ao efetuar o trancamento do curso o aluno pode ou não voltar para o mesmo.

Portanto, ao se tratar do assunto, diversos autores adotam como definição de evasão como a interrupção da jornada educacional de um estudante, antes da sua conclusão (Baggi; Lopes, 2011). Para este presente estudo, as nomenclaturas adotadas para rotular a situação dos estudantes serão de acordo com o padrão presente na base de dados do sistema acadêmico da Instituição de Ensino Superior (IE) UTFPR, em que:

- **Desistente:** discente que efetuou o cancelamento da sua matrícula;
- **Formado:** discente que concluiu a sua graduação e obteve o seu diploma;
- **Regular:** discente regularmente matriculado e com a sua graduação em andamento;
- **Trancado:** discente que suspendeu temporariamente sua matrícula;
- **Transferido:** discente que efetuou a transferência de curso e/ou campus.

2.1.2 Causas e Fatores que Influenciam

Muitos estudantes enfrentam dificuldades de adaptação à vida universitária, seja por questões pessoais ou pela incompatibilidade da IE com o mercado de trabalho. Não é possível elencar uma única razão para a ocorrência do abando do curso, mas sim uma série de fatores que se combinam, podendo até fazer com que estudantes se vejam forçados a deixar os estudos. Além disso, a falta de informação sobre os cursos e a descoberta de novos interesses também contribuem para a desistência. Jr, Ostermann e Rezende (2012) revisam em seus estudos a evasão nos cursos de graduação em Física, com base na sociologia de Pierre Bourdieu[1], onde aborda que os desafios enfrentados pelos discentes de Física são muitas vezes relacionados a questões sociais, como a falta de recursos financeiros e a falta de apoio familiar.

No relatório do MINISTÉRIO DA EDUCAÇÃO (1996), os autores definiram como principais fatores que fazem com que um discente desista de sua graduação aqueles relacionados: i) ao estudante, ii) ao curso ou a instituição e iii) fatores socioculturais e econômicos. As características individuais do estudante podem ser elencadas como habilidades de estudo, personalidade, formação escolar anterior e a escolha precoce da profissão. Já os fatores ligados as questões institucionais, são pontuados os currículos desatualizados que, de alguma maneira, frustram as expectativas do aluno, falta de comprometimento dos docentes e a falta de programa de apoio aos estudantes com dificuldades. Fatores socioeconômicos como mercado de trabalho, o reconhecimento social da carreira escolhida e as condições econômicas específicas podem influenciar diretamente as decisões dos alunos.

Biazus (2004) analisa diversos estudos e pesquisas em seu trabalho e constatou, com base nos dados analisados, que a maioria dos alunos abandonaram sua graduação devido a

[1] Pierre Bourdieu (1930-2002) foi um importante sociólogo francês com estudos relevantes no mundo todo, principalmente na área da educação.

questões pessoais, sendo a necessidade de trabalhar enquanto universitário a principal delas. Mostra também que embora a causa da evasão seja de natureza pessoal, se evidencia uma responsabilidade institucional devido à falta de programas de apoio para estudantes em situação de carência.

A deficiência na educação básica dos estudantes, especialmente aqueles provenientes de escolas públicas, representa um dos principais obstáculos para sua progressão acadêmica. Com uma base educacional frágil, os alunos enfrentam desafios significativos para acompanhar a complexidade dos conteúdos apresentados no ensino superior, impactando negativamente em sua autoestima (Cunha; Morosini, 2013).

A percepção de baixas remunerações e a falta de valorização em determinadas profissões pode levar os estudantes a reconsiderarem suas escolhas de curso. Barlem *et al.* (2012) ao analisarem a evasão nos cursos de graduação em enfermagem indicam também que a influência dos colegas desempenha um papel significativo no processo de evasão. Quando os alunos expressam descontentamento por não estarem em seus cursos de primeira escolha, cria-se um clima onde os sentimentos de frustração e insatisfação se intensificam. Essas emoções compartilhadas entre os colegas podem aumentar a pressão sobre os indivíduos, tornando a decisão pela evasão mais provável. A integração do ambiente social é essencial para a consolidação da identidade profissional, além de ser uma importante rede de apoio para lidar com as dificuldades e os desafios enfrentados ao longo da vida universitária.

Ao se analisar todas as possíveis causas e fatores anteriormente apresentados, é importante olhar para os recém ingressados na universidade com mais cautela. Estes merecem atenção especial devido aos desafios enfrentados durante o primeiro ano, e é crucial considerar os diversos ambientes que os estudantes frequentam durante essa fase (Matta; Lebrão; Heleno, 2017).

2.2 Aprendizado de Máquina (AM)

AM é um sub-campo da IA que se concentra no desenvolvimento de algoritmos e técnicas que fazem com que os computadores absorvam informações presentes nos dados e tomem as suas próprias decisões (Mitchell, 1997). Em outras palavras, é um campo que explora técnicas computacionais voltadas para a obtenção automática de novos conhecimentos, habilidades e maneiras de estruturar o conhecimento prévio.

Em Géron (2019, p.4), é definido como “a ciência (e a arte) da programação de computadores para que eles possam aprender com os dados”. O processo de aprendizagem envolve a construção de modelos matemáticos, amplamente empregados na realização de modelagens preditivas onde, após assimilar os relacionamentos que estão ocorrendo entre os dados, o algoritmo procura realizar previsões precisas e úteis ao entrarem novas amostras (Klosterman, 2020).

É um campo multidisciplinar, incorporando princípios e técnicas de diversas áreas como: Cálculo, Estatística, Computação, Estrutura de dados, entre outros. Essa interdisciplinaridade é o que permite o desenvolvimento de algoritmos eficazes, capazes de entender e manipular grandes conjuntos de dados. Tarefas de AM, como classificação, agrupamento de dados, e previsão de séries temporais, exemplificam a amplitude de domínios abrangidos (Cerri; Carvalho, 2017).

O AM tem uma ampla gama de aplicações em diversos setores, como a medicina (KOENIGKAM SANTOS *et al.*, 2019) e a agricultura (Souza *et al.*, 2016). À medida que a quantidade de dados disponíveis continua a crescer e a capacidade computacional aumenta, o AM tem se tornado cada vez mais poderoso e onipresente, impulsionando inovações e transformações em muitos aspectos da vida cotidiana. As máquinas não se limitam mais a executar apenas tarefas físicas, mas também estão desempenhando funções cognitivas (Ludermir, 2021). Além disso, diversos autores pontuam três tipos principais de AM:

- **Aprendizado Supervisionado:** para cada exemplo contido no conjunto de dados (*dataset*) é apresentada a resposta desejada (rótulo), informando à qual classe/categoria pertence a instância consultada. Chamamos de classificação as tarefas preditivas que manipulam valores categóricos, e de regressão os problemas cujos valores são representados em uma escala contínua. É a tarefa mais comum em AM, e a qual será aprofundada adiante;

- **Aprendizado Não-Supervisionado:** neste tipo de tarefa preditiva o algoritmo tenta aprender sobre dados puramente descritivos, sem a existência de rótulos, agrupando as amostras pelas similaridades dos seus atributos. Os algoritmos não-supervisionados são comumente chamados de *clustering*, e utilizam um critério de semelhança para realizar o agrupamento, como distância ou densidade, com o objetivo de identificar estruturas subjacentes nos dados;

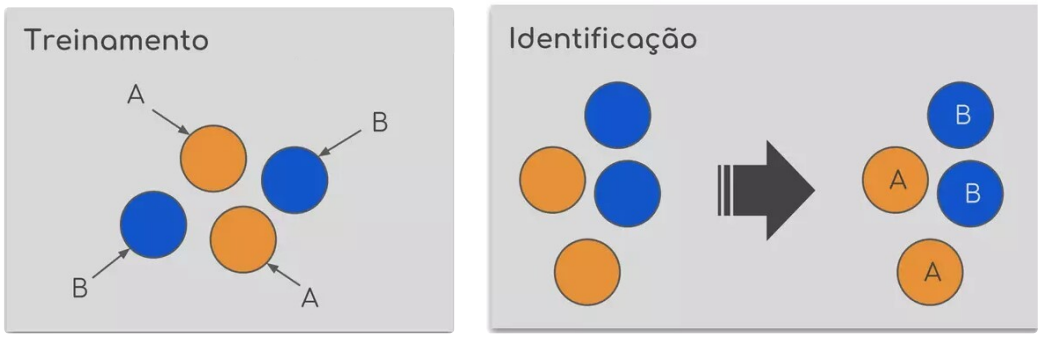
- **Aprendizado por Reforço**: neste tipo de tarefa o treinamento é feito com base em punições e recompensas. O algoritmo busca tomar sua decisão procurando obter o maior número de recompensas possíveis (Sutton; Barto, 2018), muito utilizado na robótica e em jogos.

Uma outra categoria existente do AM é o Aprendizado Semi-Supervisionado, onde são utilizados tanto dados rotulados quanto não rotulados. É um tipo de tarefa comumente utilizado nos casos em que não é possível se obter todos os rótulos, podendo ser empregado em tarefas de classificação ou agrupamento, como na classificação de textos, economizando esforço humano na rotulação manual de documentos extensos (Matsubara, 2004).

2.2.1 Aprendizado Supervisionado

Define-se uma tarefa de aprendizado supervisionado aquela em que o algoritmo utiliza rótulos para realizar o seu treinamento, objetivando prever futuras amostras que ainda não estão classificadas (Han; Pei; Tong, 2022). Nesse processo, o algoritmo é treinado com um conjunto de dados onde as entradas são associadas a saídas conhecidas, conforme exemplificado na Figura 1. A partir dessas associações, o algoritmo aprende a mapear entradas similares para as saídas corretas, permitindo que ele generalize o conhecimento adquirido.

Figura 1 – Ilustração do funcionamento do Aprendizado Supervisionado



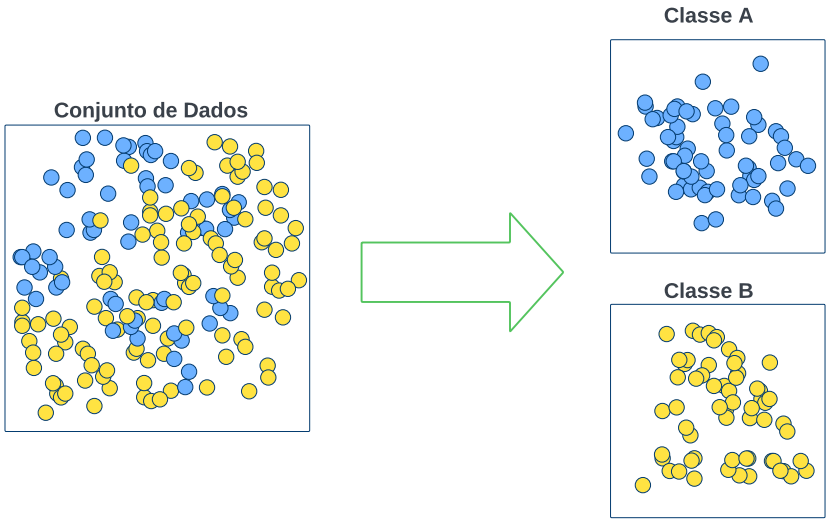
Fonte: Carvalho (2018).

Os rótulos podem ser compreendidos de duas formas distintas na análise preditiva: uma delas se relaciona com a classificação, onde o objetivo é prever categorias ou classes, enquanto a outra se refere à regressão, que visa prever valores numéricos (Silva; Peres; Boscarioli, 2017).

Os problemas de classificação envolvem a atribuição de vetores de entrada a uma das N classes disponíveis, onde cada exemplo é categorizado de forma precisa em uma única classe abrangendo todo o espaço de saída possível. No entanto, essa abordagem pode ser limitada ao lidar com uma amostra que pertence parcialmente a mais de uma classe (Marsland, 2011). A inclusão de muitos atributos pode prolongar o tempo de treinamento do classificador, além de ocasionar *overfitting*.[2] A tarefa de classificação pode ainda ser dividida em duas categorias: classificação binária, onde existem apenas duas categorias ($c = 2$) possíveis para rotular os dados; ou classificação multiclasse, onde há um número $c > 2$ de classes distintas (Silva; Peres; Boscarioli, 2017). A Figura 2 ilustra a classificação dos exemplos presentes no conjunto de dados em duas classes distintas, como exemplo de uma tarefa de classificação binária.

[2] O chamado *overfitting* é quando um modelo exibe uma escassa habilidade de generalização, indicando que a sua adaptação aos dados é excessiva. Isso acontece quando o modelo decorou ou se especializou nos dados de treinamento. Por outro lado, quando o modelo apresenta uma capacidade preditiva limitada para os dados de treinamento, pode indicar subajuste, chamado de *underfitting*, ocorrendo quando os dados de treinamento disponíveis são pouco representativos ou quando o modelo utilizado é relativamente simples e não consegue capturar os padrões presentes nos dados (Faceli *et al.*, 2021).

Figura 2 – Exemplo de uma tarefa de classificação binária



Fonte: Autoria Própria (2024).

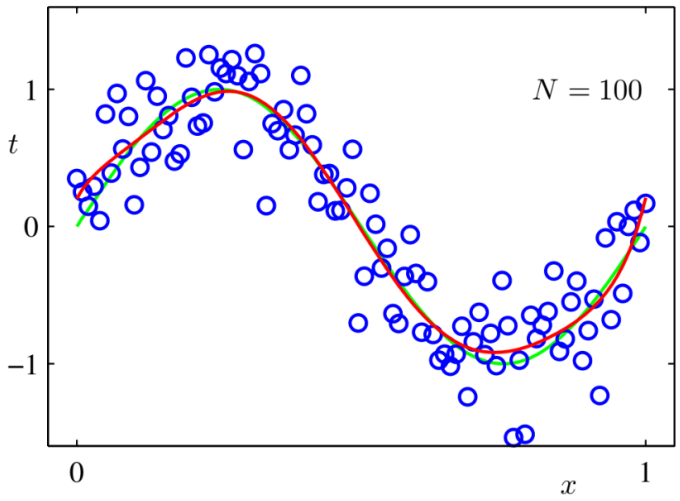
Algo que pode ocorrer em determinado conjunto de dados é a presença de classes desbalanceadas, sendo mais difícil obter informações associadas a determinados grupos. Essa discrepância na quantidade de exemplos entre as classes no conjunto de treinamento pode representar um desafio significativo para os algoritmos tradicionais, os quais podem ter dificuldade em efetuar suas predições. Isso se deve em parte ao fato de que os algoritmos tradicionais tendem a considerar igualmente importantes diferentes tipos de erros, com a suposição de que as distribuições são relativamente equilibradas (Castro; Braga, 2011).

Já uma tarefa de regressão é definida como uma técnica estatística que visa prever o valor de um ou mais rótulos contínuos, com base em um conjunto de variáveis de entrada. Essencialmente, a regressão busca encontrar a relação matemática que melhor descreve o comportamento dos dados observados. Um exemplo deste problema é a curva polinomial, onde busca-se ajustá-la representando esta relação entre as variáveis, por meio de diversas classes de funções chamadas de modelos de Regressão Linear. Outro exemplo seria a predição do rendimento em um processo de fabricação química, onde as entradas consistem nas concentrações dos reagentes, na temperatura e na pressão. A regressão pode ser expandida ao incluir uma ampla gama de funções não lineares das variáveis de entrada, chamadas de funções básicas. Elas permitem a criação de modelos flexíveis com a combinação linear dessas funções, mantendo as propriedades analíticas ao mesmo tempo em que permitem a não linearidade em relação as variáveis de entrada (Bishop, 2006).

A Figura 3 traz o exemplo do ajuste de curva polinomial, ilustrando como modelos de regressão funcionam. Neste caso, o objetivo é encontrar uma função polinomial que melhor se ajusta a um conjunto de dados observacionais. Ao ajustar uma curva polinomial aos dados, é possível capturar a relação real entre as variáveis, minimizando a diferença entre os valores previstos pelo modelo e os valores observados. Este processo envolve a determinação dos

coeficientes do polinômio que melhor descrevem a tendência dos dados, considerando o grau do polinômio como um parâmetro importante para evitar tanto o *underfitting* quanto o *overfitting*.

Figura 3 – Exemplo de Regressão



Fonte: Bishop (2006, p. 9).

2.2.2 Algoritmos de Classificação

Dentro da área do AM, diversos tipos algoritmos podem ser empregados nos problemas de classificação, cada um com características próprias que os tornam úteis para diferentes tipos de dados e problemas. Ainda podemos separar estes algoritmos de acordo com a quantidade de informação disponível para explicar as predições realizadas.

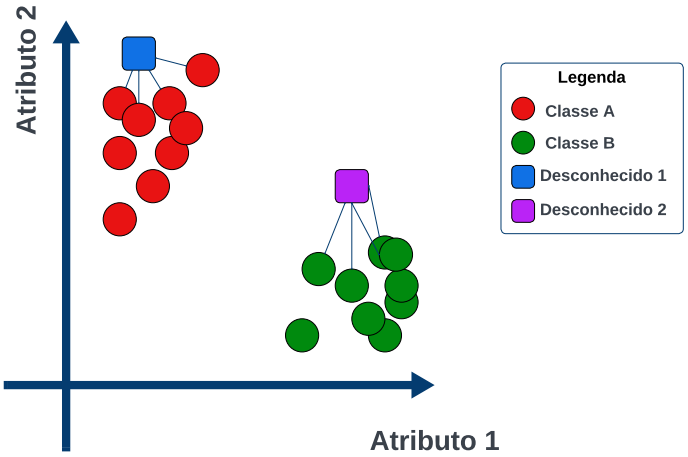
Os chamados de modelos **caixa preta** (*black box*) são aqueles considerados complexos e difíceis de interpretar, onde o seu funcionamento interno não é facilmente compreendido ou explicado diretamente pelo observador. Esses modelos geralmente oferecem um alto poder preditivo, mas não fornecem uma explicação clara sobre como as decisões são tomadas. Exemplos típicos de modelos caixa preta incluem as Redes Neurais Artificiais (RNA) e as Máquinas de Vetores de Suporte, do inglês *Support Vector Machine* (*SVM*). Por outro lado, os modelos intrinsecamente interpretáveis, também chamados de modelo **caixa branca** (*white box*), são mais simples e contém em suas estruturas e/ou parâmetros a explicação que permite entender como as decisões são tomadas (Pedulla, 2022). Exemplos de modelos caixa branca incluem Regressão Linear, Regressão Logística e a *DT*, onde as relações entre as variáveis de entrada e as saídas são claramente definidas.

Para este trabalho em específico, é crucial e desejável obter uma explicação do modelo, que possibilite o entendimento de quais características o leva a identificar se um aluno irá ou não evadir. Para isso, dentre os principais algoritmos de caixa branca existentes, serão utilizados:

- **K-vizinhos mais próximos, do inglês *K-Nearest Neighbors* (*KNN*):** é uma das técnica de aprendizado mais básicas disponíveis, se resumindo a dois elementos essen-

ciais: a noção de proximidade e a suposição de que os pontos próximos uns dos outros são semelhantes. Sua estimativa para cada ponto novo é determinada exclusivamente pelos pontos mais próximos a ele, calculando a distância do exemplo de consulta para todo o conjunto de dados, e verificando o rótulo dos K mais próximos (Grus, 2016). A classe de maior frequência entre esses vizinhos é retornada como a predição. A Figura 4 exemplifica esse processo, mostrando dois indivíduos desconhecidos sendo classificados pelos quatro vizinhos mais próximos.

Figura 4 – Exemplo de funcionamento do KNN



Fonte: Autoria Própria (2024).

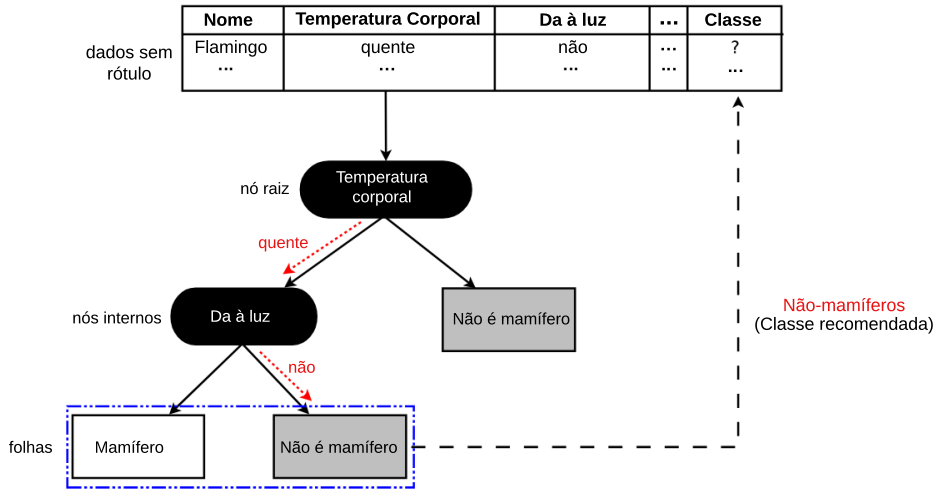
- ***Naive Bayes* (*NB*):** fundamentado no Teorema de Bayes, parte da premissa de que os atributos têm influência independente sobre a classe. Assim, ele calcula a probabilidade de um exemplo pertencer a uma classe com base na probabilidade dos valores dos atributos. Apesar de sua simplicidade, essa abordagem frequentemente produz resultados satisfatórios (McCallum; Nigam *et al.*, 1998);

- ***Decision Tree* (*DT*):** é um algoritmo que divide o conjunto de dados em subconjuntos com base nos valores dos atributos, selecionados recursivamente com base em algum critério probabilístico, como o Índice de Gini[3] por exemplo, ou a Entropia dos atributos descritivos (Monard; Baranauskas, 2003). A Figura 5 exemplifica uma *DT*, que tem como objetivo rotular se o animal é um mamífero ou não, ilustrando como cada nó realiza essa avaliação com base em critérios específicos para determinar a sua classe;

- ***Random Forest* (*RF*):** algoritmo que faz analogia a uma floresta por ser a composição de várias *DT*s, reduzindo o *overfitting* e melhorando a generalização através da combinação das previsões de cada árvore. Cada modelo (árvore) é treinado com uma

[3] O Índice de Gini é uma medida de desigualdade que quantifica uma distribuição. Variando de 0 a 1, onde 0 representa igualdade perfeita (todas as amostras pertencem a mesma classe) e 1 a desigualdade total.

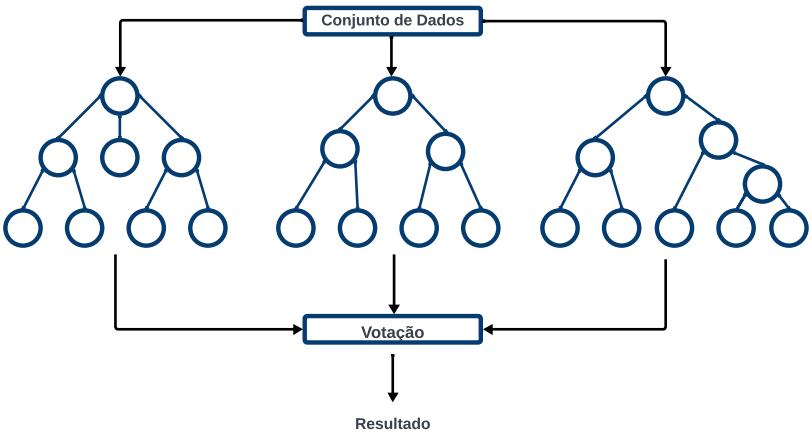
Figura 5 – Exemplo de uma *DT*



Fonte: Adaptado de Mantovani *et al.* (2024).

subamostra aleatória do conjunto de dados, geradas através do método *bootstrap*[4] (Breiman, 2001). A predição é feita pelo voto majoritário, ou seja, pelo consenso entre as árvores, onde a classe final atribuída a uma amostra é aquela que recebeu a maioria dos votos entre todas as árvores do conjunto, com uma estrutura exemplificada pela Figura 6:

Figura 6 – Exemplo de um *RF*



Fonte: Autoria Própria (2024).

2.2.3 Processo de Avaliação de Modelos

Para garantir a confiabilidade e a eficácia dos modelos de AM diversos procedimentos e métricas podem ser adotados para a avaliação dos mesmos. A divisão dos dados em conjuntos de treino e teste permite avaliar o desempenho do modelo e sua capacidade de generalização

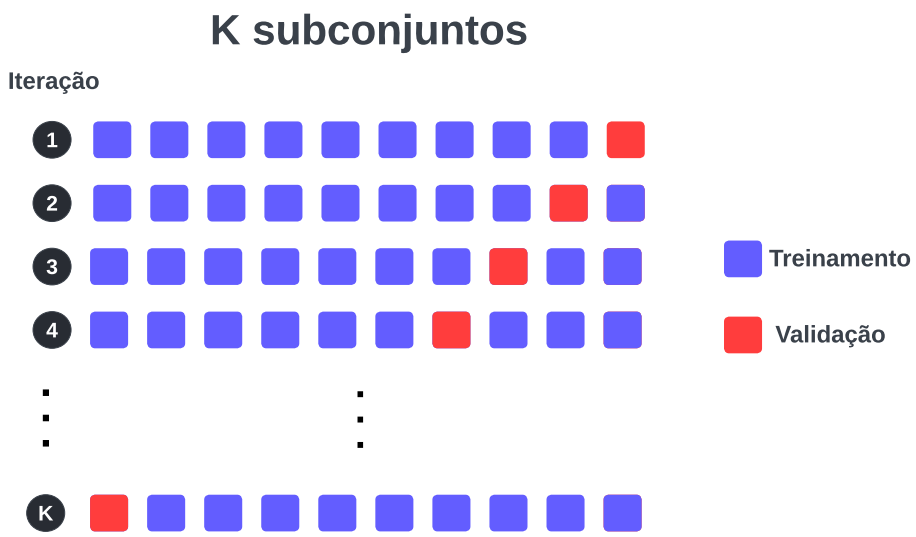
[4] *Bootstrap* é uma técnica que envolve a reamostragem dos dados de forma aleatória (Breiman, 1996).

para novos dados. Na divisão do conjunto de dados, o conjunto de treino é utilizado para treinamento do algoritmo durante o processo de aprendizado, comparando as predições obtidas com os valores esperados dos rótulos, e assim permitindo o ajuste dos parâmetros durante a indução dos modelos.

Já o conjunto de teste é reservado para avaliar o desempenho do modelo induzido em dados desconhecidos, não usados para treinamento, simulando condições reais de uso. Uma abordagem para realizar essa divisão é a validação cruzada, que pode ser implementada de forma simples ou estratificada.

O método de validação cruzada envolve a avaliação iterativa do modelo em várias partições dos dados, onde cada iteração é uma série distinta de avaliações. Ao concluir todas as iterações, o desempenho do modelo é determinado pela média das avaliações realizadas (Amaral, 2016). Em cada iteração, são criados novos conjuntos de treinamento e teste, onde múltiplas repetições são realizadas para reduzir a variação nos resultados, permitindo que todos os dados sejam utilizados tanto para treinamento quanto para teste (Castro; Ferrari, 2016). Na Figura 7 é ilustrado o seu procedimento, onde em cada iteração o conjunto de dados é dividido em partições, representadas em azul e vermelho. As partições azuis são reservadas para o treinamento dos modelos, enquanto as partições vermelhas são utilizadas para testar o desempenho dos modelos já treinados.

Figura 7 – Exemplo gráfico de Validação Cruzada



Fonte: Adaptado de Couto (2013).

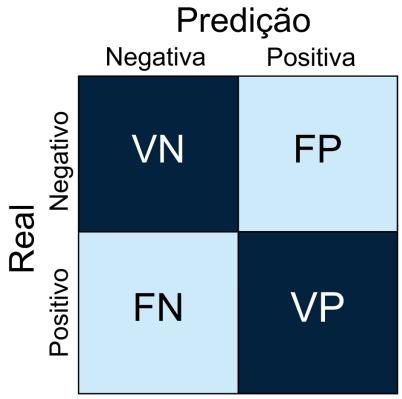
A avaliação final dos modelos preditivos pode ser conduzida de duas maneiras: i) através da análise estatística dos resultados das K avaliações, incluindo medidas como média, desvio padrão e intervalo de confiança; ii) combinando o desempenho dos K modelos produzidos e calculando a média dessa combinação em relação ao número total de exemplos no conjunto de dados original (Silva; Peres; Boscarioli, 2017).

A validação cruzada estratificada é uma variante que garante uma distribuição equilibrada das classes nas divisões, assegurando que tanto o conjunto de treinamento quanto o de teste mantenham proporções consistentes de cada classe. Isso reduz a variância nos resultados e proporciona uma avaliação mais robusta do modelo (Castro; Ferrari, 2016). Em casos com desbalanceamento de classes, a sua utilização se torna ainda mais crucial, pois permite uma avaliação mais precisa do modelo, mitigando os efeitos negativos do desbalanceamento e garantindo uma representação adequada de todas as classes durante o processo de validação.

2.2.4 Medidas de Avaliação de Modelos

Em cada uma das etapas da validação cruzada é preciso computar métricas de desempenho, para ponderar o quão boas são as predições ou não. No geral, as métricas são obtidas por meio de estatísticas aplicadas sobre a Matriz de Confusão. O uso da Matriz de Confusão representa um passo crucial no processo de avaliação dos modelos de classificação binária, fornecendo uma representação tabular das previsões em relação aos verdadeiros resultados. Dividida em quatro quadrantes, a matriz mostra os Verdadeiros Positivos (VP), Falso-Positivos (FP), Verdadeiros Negativos (VN), e Falso-Negativos (FN), sendo as duas primeiras as respostas corretas do modelo. Esses valores serão utilizados nos cálculos da Acurácia, Precisão, *Recall* e *F1-Score* (Liang, 2022). A Figura 8 mostra uma Matriz de Confusão para um problema de uma classificação binária. Quanto maiores os valores na diagonal principal, maior a precisão do modelo.

Figura 8 – Exemplo de uma Matriz de Confusão



Fonte: Autoria Própria (2024).

A medida **Acurácia Simples** é a métrica mais simples a ser computada. Ela representa a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. Seu cálculo pode ser representado como:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

No entanto, é importante considerar o contexto específico do problema, pois em conjuntos de dados desbalanceados a acurácia pode ser enganosa, favorecendo classes majoritárias. Esse problema surge porque, de acordo com sua a formulação matemática, os casos de VN podem ofuscar os casos de VP, podendo levar a uma interpretação equivocada do desempenho do modelo. Para isso, a **Acurácia Balanceada, do inglês *Balanced Accuracy* (*BAC*)** garante uma avaliação mais precisa em relação as diferentes classes, não sendo influenciada pelo seu desequilíbrio, uma vez que seus cálculos se baseiam nas taxas de VP e VN (Braga:2022). A sua fórmula matemática pode ser representada como:

$$BAC = \frac{1}{2}\left(\frac{VP}{VP+FN} + \frac{VN}{VN+FP}\right) \tag{2}$$

O *Recall* é uma métrica utilizada para avaliar a capacidade de um modelo em identificar corretamente todos os exemplos positivos de uma classe. Em outras palavras, mede a proporção de exemplos positivos que foram corretamente identificados pelo modelo em relação ao total de exemplos positivos existentes (Powers, 2020). Sua fórmula pode ser representada como:

$$Recall = \frac{VP}{VP+FN} \tag{3}$$

Já a precisão (*precision*), irá medir a proporção de predições positivas que realmente são, em relação ao total de exemplos previstos como positivos. Ela é calculada por:

$$Precisão = \frac{VP}{VP+FP} \tag{4}$$

Por fim, o *F1-score* combina o *Recall* e a Precisão em uma única métrica, levando em consideração tanto os FP quanto os FN. Isso é especialmente útil quando há um desequilíbrio entre as classes no conjunto de dados (Powers, 2020). Sua fórmula é dada por:

$$F1 - Score = 2 * \frac{Precisão * Recall}{Precisão + Recall} \tag{5}$$

Essas métricas desempenham um papel essencial na avaliação precisa dos resultados da previsão de evasão estudantil, fornecendo uma compreensão detalhada de como o modelo se sai em termos de precisão, sua habilidade de identificar casos positivos e sua capacidade de se adaptar a novos dados. É crucial ter um entendimento adequado dessas métricas para embasar decisões informadas sobre quais modelos escolher, como refiná-los e como aplicá-los de forma eficaz na prática.

2.3 Considerações Finais

Ao longo deste capítulo, foram apresentados os conceitos teóricos fundamentais que servirão como base para a compreensão e desenvolvimento do trabalho. Desde a definição dos problemas abordados até a exposição das metodologias e métricas relevantes, cada seção contribui para uma compreensão sólida e abrangente do contexto em que a pesquisa se insere. Com essa base estabelecida, o próximo passo será a aplicação prática desses conhecimentos na análise e resolução dos desafios propostos.

3 TRABALHOS RELACIONADOS

Em meio ao crescimento de estudos envolvendo o *EDM*, a evasão estudantil nas Instituições de Ensino Superior (IEs) vem sendo cada vez mais estudada e debatida no meio acadêmico. Pesquisadores e educadores têm dedicado a sua atenção para compreender os motivos que levam os alunos a abandonarem os estudos, na tentativa de minimizá-lo e preveni-lo. O AM entra como uma excelente ferramenta para identificar padrões e características dos alunos que desistem de seus cursos, proporcionando análises valiosas para o desenvolvimento de medidas personalizadas e eficazes. O objetivo deste capítulo é apresentar o estado da arte e as soluções que vêm sendo desenvolvidas na literatura relacionada, fornecendo uma análise criteriosa de estudos relevantes que investigam uma variedade de metodologias e aplicações.

Durante a revisão bibliográfica, foram analisados 20 artigos acadêmicos relevantes no campo em questão. Nas seções seguintes, serão discutidos os artigos mais pertinentes, divididos entre os internacionais e os produzidos aqui no Brasil. Através dessa análise, será possível destacar as contribuições mais significativas encontradas na literatura atual. Além disso, abaixo é apresentado na Tabela 1 todos os artigos analisados, acompanhados de suas principais características, fornecendo uma visão abrangente do panorama geral. Na tabela, são listados cada um destes trabalhos, em ordem: identificador da citação original do trabalho; o tipo de tarefa preditiva realizada; os algoritmos de AM explorados pelos autores; as métricas de desempenho utilizadas para avaliar os modelos preditivos; e características gerais que destacam o trabalho dos demais estudos da literatura.

3.1 Uso de AM para predição da evasão estudantil

3.1.1 Trabalhos Internacionais

Solis *et al.* (2018) realizaram uma análise detalhada da predição da evasão de estudantes universitários no Instituto Tecnológico de Costa Rica. Nos experimentos, fizeram uso de quatro algoritmos de AM (*RF*, Redes Neurais, *SVM* e Regressão Logística) e quatro perspectivas distintas para definir os arquivos de dados, que incluíam: i) considerar todos os registros de matrícula, classificando como não evadido alunos ativos ou graduados; ii) excluir alunos ativos, considerando como não evadido apenas aqueles que já concluíram o curso; iii) usar apenas o último semestre antes da evasão para evadidos ou um semestre aleatório para não evadidos; iv) mesma definição de evadidos e não evadidos da segunda perspectiva, mas utilizando um único semestre por aluno, para eliminar o ruído de múltiplos semestres. Os resultados revelaram que o algoritmo *RF*, alcançando uma sensibilidade de 93% na previsão de evasões da quarta perspectiva. Este estudo é uma base importante para o presente trabalho, ressaltando a importância de considerar diversas perspectivas na definição dos dados.

Tabela 1 – Trabalhos Relacionados sobre Evasão Estudantil Usando Aprendizado de Máquina

Identificador	Tarefa Preditiva	Algoritmos de AM utilizados	Métricas de Avaliação Utilizadas	Características
Manhães, Cruz e Zimbrão (2014)	Classificação Binária	*DT*, *NB*, MLP, *SVM*	Acurácia	Artigo mais antigo encontrado e revisado. Apresenta uma arquitetura de três camadas onde se apresenta, aplica e armazena os dados.
Paz e Cazella (2017)	Classificação Binária	*DT* (J48)	Acurácia	Analisou o impacto de bolsas de estudo fornecidas pela IE, além de considerar a localização dos alunos em relação ao seu município de origem.
Solis *et al.* (2018)	Classificação Binária	*RF*, RNA, RL e *SVM*	Especificidade, *Recall*, *Kappa* e Precisão	Uso de quatro perspectivas distintas para dividir o conjunto de dados.
Rodrigues (2018)	Classificação Binária	*DT*, *KNN*, RNA, RL e *SVM*	Acurácia, Precisão, *Recall* e AUC	Análise comparativa dos classificadores de AM para prever a evasão de alunos em cursos Educação a Distância (EaD).
Saraiva *et al.* (2019)	Classificação Binária	*KNN*, RNA, *RF* e *SVM*	Acurácia e Precisão	Aplicação de técnicas de visualização de dados com o software *Tableau* para explorar e entender os dados acadêmicos e socioeconômicos dos estudantes.
Ahmed e Khan (2019)	Classificação Binária	*DT*, *KNN*, *NB*, RN, RL, *RF* e *SVM*	Acurácia, Precisão, *Recall*, *F1-Score* e AUC	Coletando dados de 480 estudantes através de um questionário, e explorou utilizando técnicas de Mineração de Dados.
FIlho, Vinuto e Leal (2020)	Classificação Binária	*DT*, *KNN*, *NB* e *SVM*	Acurácia, Precisão, Revocação, Especificidade e *F1-Score*	Utilização do *Gradient Boosting* (combinação de múltiplos modelos fracos para melhorar a previsão).
Teodoro e Kappel (2020)	Classificação Binária	*DT*, *KNN*, *NB*, RN e *RF*	Acurácia, Precisão, *Recall*, *F1-Score*, ROC e AUC	Análise detalhada de padrões críticos que influenciam a evasão.

Continua na próxima página

Identificador	**Tarefa Preditiva**	**Algoritmos de AM utilizados**	**Métricas de Avaliação Utilizadas**	**Características**
Sonnenstrahl, Bernardi e Pertile (2021)	Classificação Binária	*DT* (J48), *MLP NB*, RN e IDK	Acurácia, Precisão, *Recall* e *F1-Score*	Utilizaram dados de interações com os recursos online e atividades entregues para prever a evasão.
Kabathova e Drlik (2021)	Classificação Binária	*DT*, *NB*, RN, *RF*, RL e *SVM*	Acurácia, Precisão, *Recall* e *F1-Score*	Comparação de seis algoritmos de AM, com destaque para a importância do ajuste dos hiperparâmetros.
Colpo, Primo e Aguiar (2021)	Classificação Binária	*DT* e *RF*	Acurácia, Precisão e *Recall*	Introduziu dois tipos de representações: MRE-URF (múltiplos registros para evadidos e único registro para formados) e a URE-MRF (único registro para evadidos e múltiplos registros para formados).
Santos (2022)	Classificação Binária	*DT*, *LightGBM* e *SVM*	Acurácia, Precisão, *Recall* e *F1-Score*	Utilização de questionários online, além dos dados do sistema acadêmico.
Júnior *et al.* (2022)	Classificação Binária	*NB*, *KNN* e *RF*	Acurácia, Precisão, *Recall* e *F1-Score*	Investiga a evasão escolar na educação básica.
Barros (2022)	Classificação Binária	*DT*, *Extra Trees*, *KNN*, *NB*, *RF* e *SVM*	Acurácia, Precisão, *Recall*, *BAC* e *F1-Score*	Utilização de técnicas como *Gradient Boosting* e *XGBoost*.
Silva (2022)	Classificação Multiclasse	*DT*, *NB*, RN e RL	Acurácia, Especificidade, *Recall* e *F1-Score*	Busca classificar a situação dos estudantes, considerando diferentes categorias: Cursando, Desvinculado, Trancado, Transferido e Formado.
Martins *et al.* (2023)	Classificação Binária	*DT*, *KNN*, *RF* e RL	Acurácia, Precisão *Recall* e *F1-Score*	Enfoque na predição da evasão tardia.
Vossen *et al.* (2023)	Classificação Binária	*AutoGluon*	-	Utilização de técnicas de *AutoML*.
Donna (2023)	Classificação Multiclasse	*DT*, *MLP* e *RF*	Acurácia, Precisão *Recall* e *F1-Score*	Busca rotular a situação dos alunos dentre três classes: Cancelado, Formado e Matriculado.

Continua na próxima página

Identificador	**Tarefa Preditiva**	**Algoritmos de AM utilizados**	**Métricas de Avaliação Utilizadas**	**Características**
BRAND C. *et al.* (2024)	Classificação Binária	*K-Means*, *Isolation Forest*, dt e RL	Acurácia e Precisão	Utilização de técnicas de AM supervisionado e não supervisionado.
Ahmed (2024)	Classificação Binária	*DT*, *NB*, *KNN* e *SVM*	Acurácia, Precisão, *Recall* e *F1-Score*	Utilização da ferramenta *WEKA*.

Fonte: Autoria Própria (2024).

Ainda no âmbito internacional, Ahmed e Khan (2019) realizou um estudo de evasão em Bangladesh, um país que semelhante ao Brasil, enfrenta altas taxas de evasão no ensino superior. Os autores procuraram então antecipar alunos em risco de evasão por meio do uso de técnicas de *EDM*, onde técnicas de mineração de dados são empregadas para identificar padrões e tendências em informações educacionais. Nos experimentos realizados, dados foram coletados de 480 estudantes de várias cidades de Bangladesh por meio de um questionário enviado a ex-alunos. Esse questionário continha 28 perguntas que exploravam fatores pessoais, acadêmicos e institucionais. Os dados foram divididos em três conjuntos: um com dados pessoais, outro com dados acadêmicos e um terceiro com todos os tipos de dados (acadêmicos, pessoais e institucionais) para determinar o modelo mais eficaz. Entre os algoritmos utilizados, as Redes Neurais demonstraram o melhor desempenho, com uma precisão de $0{,}915$. Embora não tenha sido informado qual modelo de Redes Neurais (RN) foi utilizado, suponha-se que seja uma *Multi-layer Perceptron* (*MLP*), com base no contexto fornecido.

Em um estudo mais recente, Ahmed (2024) conduziu uma pesquisa com o intuito de prever a evasão de estudantes de faculdades de educação devido ao baixo desempenho acadêmico. Para isso, selecionou quatro instituições de ensino superior público no nordeste da Nigéria e incluiu todos os alunos do primeiro ano da Escola de Educação em Ciências dessas instituições. Os dados foram coletados em duas etapas distintas, abrangendo os resultados dos alunos no primeiro e segundo semestres do ano acadêmico 2022/2023, além do desempenho pré-universitário dos alunos no UTME. Utilizando a aplicação WEKA[1], desenvolveu os modelos de AM. Os resultados mostraram que a *SVM* foi o modelo mais eficaz na previsão da evasão, com uma precisão de $0{,}92$ e um *F1-Score* de $0{,}87$, sendo recomendada para a detecção precoce de alunos em risco de abandonar os estudos.

BRAND C. *et al.* (2024) conduziram um estudo para desenvolver uma aplicação web destinada a identificar a probabilidade de abandono de estudantes no Serviço Nacional de Aprendizagem (SENA) na Colômbia. Durante o estágio de modelagem, técnicas de AM supervisionado e não supervisionado foram exploradas. Em termos de avaliação e resultados, o projeto alcançou uma precisão impressionante de 91% com o modelo de *DT*. Além disso, destacou-se a importância dos fatores socioeconômicos nas taxas de evasão e propôs soluções baseadas em IA para aprimorar o bem-estar dos estudantes e estratégias de retenção.

3.1.2 Trabalhos desenvolvidos no Brasil

FIlho, Vinuto e Leal (2020) abordaram a questão crítica da evasão estudantil nos campi mais afetados do Instituto Federal do Ceará, onde as taxas eram superiores a 40%. Enqua-

[1] A aplicação WEKA é uma ferramenta de código aberto utilizada para mineração de dados e AM, desenvolvida pela Universidade de Waikato, na Nova Zelândia. Oferece a possibilidade de utilizar diversos algoritmos, realizar o pré-processamento de dados, além da visualização e avaliação dos modelos.

drando como um problema de classificação binária, utilizaram dados de cinco campus obtidos através da plataforma *IFCE em Números*, de 2015 a 2019. As variáveis selecionadas para a análise incluíam: se o aluno era ingressante no período ou não, se estava retido no curso ou não, se era cotista ou não, a etnia do aluno, o sexo, o nível de ensino (técnico, graduação ou básico), a faixa etária, e se o aluno era natural do mesmo município da instituição ou de outro município. Os dados foram pré-processados através da remoção de duplicidades, correção de termos, preenchimento de valores nulos e conversão de variáveis categóricas textuais para numéricas. Em seguida, foram aplicados quatro algoritmos de AM (*NB*, *DT*, *SVM*, e *KNN*), avaliados com base em cinco métricas: Acurácia, Precisão, Revocação, Especificidade e *F1-Score*. O melhor desempenho foi observado com *Gradient Boosting*, alcançando *F1-Score* de $0{,}87$ no campus Quixadá. Nos demais algoritmos, os valores de *F1-Score* variaram entre $0{,}58$ e $0{,}69$.

Já Teodoro e Kappel (2020) conduziram uma pesquisa focada na identificação dos padrões que indicam uma maior probabilidade de evasão entre os alunos das instituições públicas de ensino superior no Brasil. Utilizando dados do INEP, foram aplicadas cinco algoritmos de AM, destacando o *RF* como a mais eficaz com uma taxa de acerto de cerca de 80% na previsão de evasão. Os resultados revelaram que características como idade, participação em atividades extracurriculares e carga horária do curso são determinantes na evasão.

Tratando exclusivamente da evasão no EaD, Sonnenstrahl, Bernardi e Pertile (2021) contextualizaram a sua importância no cenário educacional atual e exploraram suas causas, desde a falta de interação face a face até os desafios de motivação e suporte institucional. A pesquisa abordou uma análise das interações dos estudantes no Ambiente Virtual de Aprendizagem (AVA), sendo o principal indicador para prever a evasão de estudantes em cursos EaD. Os dados continham a quantidade de cliques dos alunos em recursos online como vídeos de aula, documentos de leitura e número de atividades entregues. O algoritmo *RF* obteve a melhor acurácia no primeiro conjunto de dados, sendo $93{,}33\%$ com a visualização de tarefas se destacando como o indicador principal, seguido pelo J48, com $85{,}71\%$ com a submissão de tarefas como atributo principal.

Em sua tese de mestrado, Santos (2022) abordou técnicas de AM na Universidade Federal de São Carlos (UFSCar) para realizar a identificação dos discentes em uma possível situação de abandonarem o curso. O autor modelou o problema como uma tarefa de classificação binária. Com isso, pode executar o seu objetivo principal, sendo a criação de relatórios eletrônicos contendo dados e estatísticas relevantes, destinados a auxiliar os gestores universitários, líderes de departamentos e cursos, bem como os professores, a tomarem decisões embasadas para reduzir a evasão. Utilizando dados coletados dos sistemas acadêmicos e questionários online, pôde também identificar características chave e o contexto da evasão na instituição. Os resultados mostraram que a mineração de dados teve uma alta capacidade para prever os estudantes em risco de evasão, especialmente nos primeiros anos. A utilização de dados de todos os anos anteriores apresentou um desempenho superior em termos de precisão, atingindo o valor de $0{,}83$ e um *F1-score* de $0{,}80$ com o *LightGBM*.

Analisando a educação básica, o estudo de Júnior *et al.* (2022) aborda a aplicação de técnicas de *EDM* para prever a evasão escolar em colégios públicos de Salvador. Uma abordagem proativa foi adotada para contornar as restrições de acesso aos dados, utilizando um formulário online para coletar informações dos alunos. O estudo aponta para possíveis intervenções institucionais e pedagógicas para mitigar o problema. O *RF* teve o melhor desempenho nos testes, com acurácia de $0{,}919$ e *Recall* de $0{,}938$. No entanto, são reconhecidas as limitações, como o tamanho da amostra e a falta de acesso direto aos dados do governo, e sugere que futuras pesquisas possam explorar parcerias institucionais para ampliar o escopo e a eficácia dos modelos de previsão de evasão escolar.

O estudo de Martins *et al.* (2023), realizado na Universidade Federal de Juiz de Fora (UFJF), investiga a evasão estudantil no ensino superior, focalizando especialmente a evasão tardia, onde o aluno evadiu após o quinto período. Utilizando dados de estudantes de graduação presencial entre 2003 e 2020, abordou diferentes perspectivas da evasão estudantil, incluindo suas causas internas e externas, bem como os impactos econômicos e sociais decorrentes. Com base em metodologias robustas de pré-processamento, transformação e seleção de dados, os pesquisadores empregaram algoritmos como *DT*, *KNN*, Regressão Logística e *RF* para desenvolver modelos preditivos. O protocolo de validação estratificada e as métricas de avaliação, como acurácia, precisão, recall e *F1-score*, foram aplicados para avaliar o desempenho dos modelos. Os resultados revelaram diferenças significativas no desempenho dos algoritmos, com destaque para o *RF* como o modelo mais eficaz na previsão da evasão tardia, tanto na base geral quanto na base de finalistas, com $0{,}93$ na base geral e $0{,}94$ e $0{,}88$ de *F1-score*.

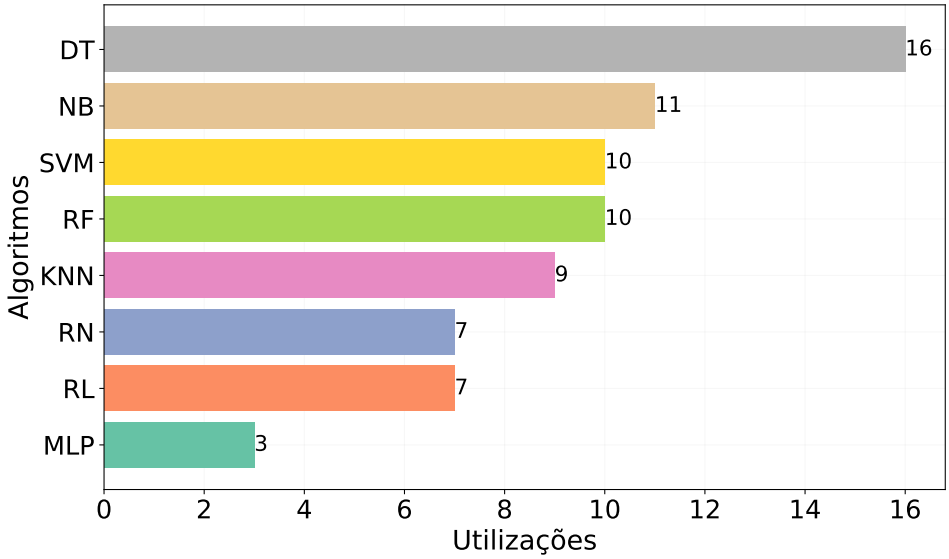
3.2 Análise da Literatura

Após a análise de diversas pesquisas relacionadas, é possível extrair algumas conclusões e identificar tendências relevantes para o desenvolvimento de novos trabalhos nesta área. A aplicação de algoritmos de AM, como *RF* e *DT*, tem se mostrado eficaz na previsão de evasões, alcançando altas taxas de acerto em diversos contextos educacionais. Os trabalhos analisados mostraram também a importância de um pré-processamento cuidadoso dos dados. Estes processos podem impactar significativamente o desempenho dos modelos de predição. Vários estudos também ressaltaram a importância de se considerar a aplicabilidade e a generalização dos modelos desenvolvidos, visto que em muitos casos as classes se encontravam desbalanceadas.

É possível identificar, em meio a uma variedade de algoritmos de AM comumente empregados, aqueles que foram mais utilizados. Dentre eles, estão o *RF*, *NB* e o *DT*, como apresentados na Figura 9, devido a suas eficácias comprovadas em diversos problemas de AM, além de serem facilmente interpretáveis.

Além disso, é possível notar uma alta utilização de algoritmos de caixa preta, como o *SVM* e as RN. Embora esses algoritmos possam oferecer um alto desempenho preditivo, sua

Figura 9 – Algoritmos mais utilizadas nos artigos revisados

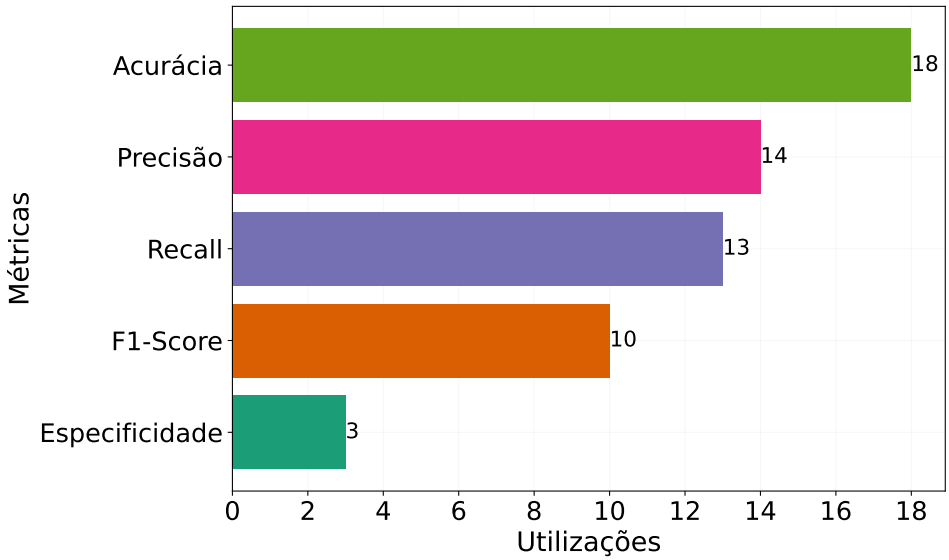


Fonte: Autoria Própria (2024).

complexidade muitas vezes dificulta a interpretação dos resultados e a compreensão dos fatores subjacentes que influenciam as suas decisões. Além disso, são altamente sensíveis aos ajustes de hiperparâmetros, podendo necessitar de uma otimização para se obter bons resultados.

Quanto as medidas de avaliação comumente empregas, estão entre as mais utilizadas a Acurácia (simples), Precisão (*Precision*), Revocação (*Recall*), Especificidade (*Specificity*) e *F1-Score*. Estas métricas fornecem uma visão abrangente do desempenho dos modelos em diferentes aspectos, como a capacidade de fazer previsões corretas, tanto positivas quanto negativas, e a confiabilidade de um modelo preditivo. A Figura 10 mostra as mais utilizadas nas pesquisas revisadas.

Figura 10 – Métricas mais utilizadas nos artigos revisados



Fonte: Autoria Própria (2024).

3.3 Considerações Finais

Os estudos revisados nesta análise demonstraram uma predominância de abordagens de classificação binária na predição da evasão estudantil. Essa escolha pode ser atribuída a natureza direta e objetiva da tarefa, onde o objetivo é simplesmente prever se um aluno irá evadir ou não.

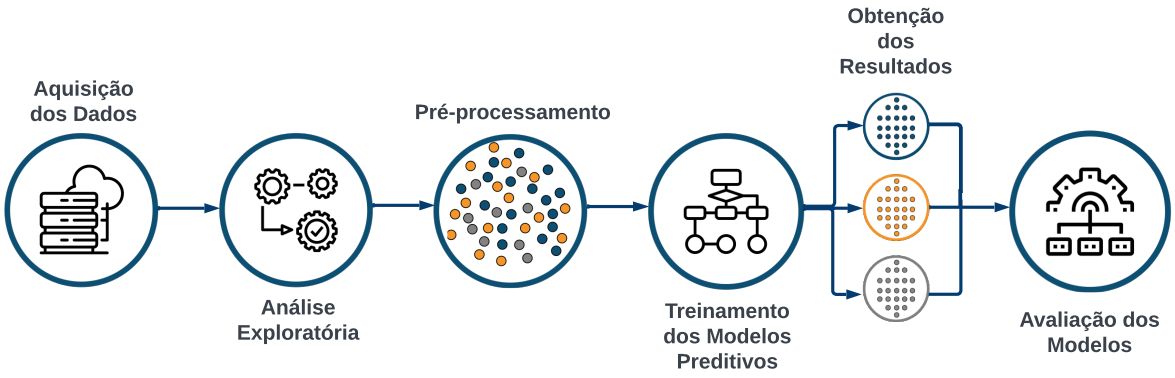
A análise da divisão do conjunto de dados em diferentes critérios, conforme explorado nos estudos revisados, demonstrou ser uma abordagem valiosa. A variedade de perspectivas adotadas, como a definição de evasão e a delimitação temporal dos dados, permitiu uma visão abrangente do fenômeno em diferentes contextos educacionais. Com base nessa análise, o presente trabalho adotará uma abordagem semelhante para dividir o conjunto de dados, considerando múltiplos critérios que possam influenciar o desempenho dos modelos de AM.

Uma área que ainda foi pouco explorada é a realização de testes estatísticos comparativos entre modelos de AM. A sua aplicação será um ponto relevante a ser explorado neste trabalho, fornecendo uma análise mais abrangente e robusta do desempenho dos algoritmos de AM na predição da evasão estudantil.

4 METODOLOGIA EXPERIMENTAL

A Figura 11 apresenta a metodologia experimental adotada no desenvolvimento deste trabalho, representando o *pipeline*[1] completo e todas as sub-etapas necessárias para a elaboração da solução de AM, indo da aquisição dos dados até a avaliação dos modelos. Cada uma das próximas seções deste capítulo descreve detalhadamente estas sub-etapas realizadas no experimento.

Figura 11 – Etapas da Metodologia



Fonte: Autoria Própria (2024).

4.1 Aquisição dos dados

A aquisição dos dados foi feita a partir do Sistema Acadêmico da UTFPR, campus de Dois Vizinhos, com aprovação formal feita ao Comitê de Ética em Pesquisa (CEP) da referida universidade, com número 63867122.1.0000.0177. Foram coletadas informações relacionadas ao desempenho acadêmico dos estudantes, como coeficiente de rendimento e número de disciplinas aprovadas e reprovadas, e caraterísticas da sua forma de ingresso na instituição. Para garantir a integridade e privacidade dos dados, todos os identificadores pessoais e informações sensíveis foram removidos. Assim, foi utilizada uma versão anonimizada dos dados nos experimentos.

A base de dados original continha um total de 44.129 registros de estudantes, abrangendo 7 cursos de graduação ativos e 1 inativo. Os cursos atualmente em funcionamento incluem o Bacharelado em Zootecnia, o Bacharelado em Agronomia, Engenharia Florestal, Bacharelado em Engenharia de Software, Licenciatura em Ciências Biológicas, Engenharia de Bioprocessos e Biotecnologia, e Licenciatura em Educação no Campo. Já o único curso inativo registrado na base é o Curso Superior de Tecnologia em Horticultura.

[1] Um *Pipeline* é uma sequência de etapas interligadas de processamento e modelagem de dados, desenvolvida para automatizar, padronizar e acelerar o processo de criação, treinamento, avaliação e implantação de modelos de AM (Géron, 2019).

Para este estudo, optou-se por utilizar apenas os dados referentes aos cursos de graduação ainda ativos, a fim de refletir a realidade atual da instituição e garantir que a análise se concentre em programas em operação, excluindo aqueles que foram descontinuados ou encerrados. Essa seleção resultou em um conjunto de 43.883 registros, onde cada linha representa um semestre cursado por um estudante. A escolha de incluir exclusivamente cursos ativos visa assegurar a relevância e aplicabilidade dos resultados, uma vez que esses cursos continuam a influenciar diretamente a experiência acadêmica dos alunos atuais e futuros.

A UTFPR possui critérios específicos para calcular o período atual do aluno. Por exemplo, para ser considerado no quarto período, o aluno não pode acumular uma carga horária superior a 16 horas semanais de aula referente a matérias dos períodos anteriores. Devido a essa regra, alguns alunos possuem múltiplos registros em um mesmo período. Todas as características extraídas do sistema acadêmico estão listadas na Tabela 2.

4.2 Tarefas Preditivas

Para a realização dos experimentos, foram definidas algumas variações do dataset original: i) contendo todos os registros extraídos do sistema acadêmico (43.883), ou ii) apenas o registro mais recente de cada aluno. Essa diferenciação visa compreender se a totalidade dos dados de cada aluno — ou seja, seu histórico completo — influencia no desempenho preditivo dos algoritmos, em comparação ao uso apenas do último estado acadêmico registrado. A exemplo, se um aluno esta no oitavo semestre de seu curso, no conjunto de dados filtrado apenas o registro correspondente a este semestre final será mantido. Em contrapartida, no conjunto de dados completo, todos os registros de semestres anteriores também estarão incluídos. Mantendo apenas os registros mais recentes de cada aluno, reduzimos o total de exemplos para apenas 5.656 registros.

A segunda variação de datasets nos experimentos envolveu a definição dos atributos alvo (classes). A tarefa de predição de evasão foi formulada como um problema de classificação binária, categorizando as instâncias em duas classes principais: “Regular” e “Desistente”. A partir dessa estrutura inicial, foram criadas cinco variações da tarefa, cada uma explorando diferentes combinações das categorias Regular e Desistente, conforme descrito abaixo:

- **Primeira Tarefa:** Classes ’Formado + Regular’ versus ’Desistente + Trancado’. Alunos na situação ’Formado’ foram classificados como ’Regular’, e os alunos com situação ’Trancado’ foram classificados como ’Desistente’;
- **Segunda Tarefa:** Classes ’Regular’ versus ’Desistente + Trancado’;
- **Terceira Tarefa:** Classes ’Formado + Regular’ versus ’Desistente’;
- **Quarta Tarefa:** Classes ’Formado’ versus ’Desistente’;

Tabela 2 – Características presentes na base de dados original

Características	
Ano de Ingresso	Calouro
Categoria Stricto Sensu	Coeficiente de Rendimento
Campus	Sede
Código	Situação
Código INEP do Curso	Data de Colação de Grau
Data de Ingresso	Data de Nascimento
Disciplinas Aprovadas	Disciplinas Consignadas
Disciplinas Matriculadas	Disciplinas Reprovadas Por Frequência
Disciplinas Reprovadas Por Nota	E-Mail
Forma de Ingresso	Idade
Gênero	Ano
Município	Município SISU
Mudou de Curso - Mesmo Campus	Mudou de Curso - Outro Campus
Nível de Ensino	Grau
Nome	Semestre
Nota ENEM Humanas	Nota ENEM Linguagem
Nota ENEM Matemática	Nota ENEM Natureza
Nota ENEM Redação	Nota Final (SISU/Vestibular)
Nota Vestibular Biologia	Nota Vestibular Filosofia e Sociologia
Nota Vestibular Física	Nota Vestibular Geografia
Nota Vestibular História	Nota Vestibular Literatura Brasileira
Nota Vestibular Língua Estrangeira Moderna	Nota Vestibular Língua Portuguesa
Nota Vestibular Matemática	Nota Vestibular Química
Número De Entradas em Outros Cursos	Número de Entradas No Curso
Ordem de Ingresso no Curso	País de Nascimento
Periodicidade	Funcionamento
Provável Jubilamento	Regime de Ensino
Retenção Parcial	Retenção Total
Semestre de Ingresso do Curso	Sede
Tipo de Cota	Total de Períodos do Curso
Total De Semestres Cursados	UF
UF SISU	Período

Fonte: Autoria Própria (2024).

- **Quinta Tarefa:** Classes ’Regular’ versus ’Desistente’.

Cada uma dessas tarefas resultou na criação de um novo *dataset*, contendo apenas a última ocorrência ou todas as ocorrências de cada aluno. Cada uma destas tarefas foi definida seguindo as duas variações de datasets, com todos ou apenas os últimos registros de um aluno. Assim, no total, são geradas 10 diferentes tabelas preditivas, como apresentado na Tabela 3.

Adicionalmente, para melhor compreender o comportamento dos alunos ao longo dos períodos acadêmicos, foram definidas tarefas adicionais gerando conjuntos de dados para diferentes intervalos de períodos, sendo esses: do 1º ao 3º, do 1º ao 5º e do 4º ao 10º períodos. Nestes cenários, alguns atributos descritivos foram removidos dos datasets, como: *período*, *ano*, *semestre* e *ano de ingresso*, permitindo uma análise sem esses dados temporais específicos.

Tabela 3 – Tarefas de classificação binária geradas para os experimentos

Tarefa	Acrônimo (Não Evadido vs Evadido)	Amostras	Não Evadido	Evadido	Parcela de Evadidos
1	Últimos Registros de Formado + Regular vs Desistente + Trancado	5508	2975	2533	0.46
2	Últimos Registros de Regular vs Desistente + Trancado	3870	1337	2533	0.65
3	Últimos Registros de Formado + Regular vs Desistente	5351	2975	2376	0.44
4	Últimos Registros de Formado vs Desistente	4014	1638	2376	0.59
5	Últimos Registros de Regular vs Desistente	3713	1337	2376	0.64
1	Completo com Formado + Regular vs Desistente + Trancado	43426	37073	6353	0.15
2	Completo com Regular vs Desistente + Trancado	41766	35413	6353	0.15
3	Completo com Formado + Regular vs Desistente	39843	37073	2770	0.07
4	Completo com Formado vs Desistente	4430	1660	2770	0.63
5	Completo com Regular vs Desistente	38183	35413	2770	0.07

Fonte: Autoria Própria (2024).

Desta forma, considerando todos os experimentos realizados, foram gerados um total de 180 variações diferentes do dataset original.

4.3 Análise Exploratória

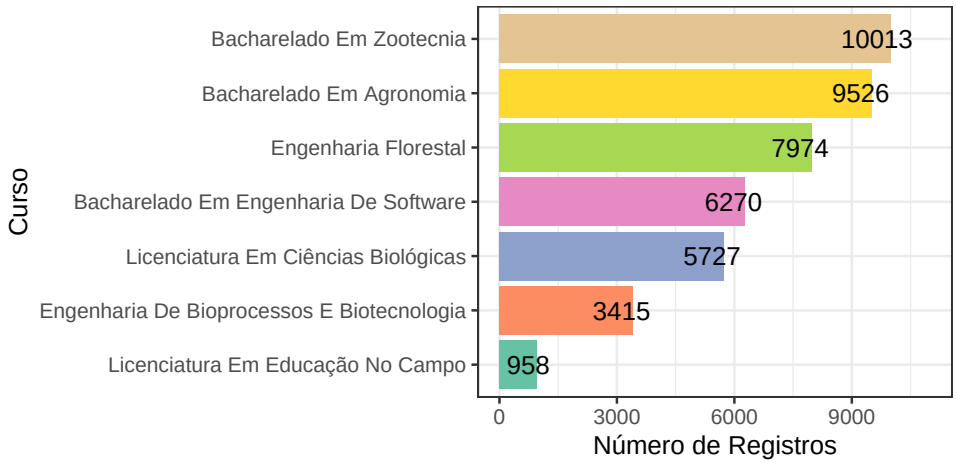
Antes de aplicar os modelos preditivos, realizou-se uma Análise Exploratória sobre os dados (*Exploratory Data Analysis - EDA*), para identificar distribuições e características intrínsecas, gerando visualizações que auxiliam na sua compreensão. Essa análise foi feita nos dois cenários: considerando todos os registros históricos e considerando apenas o último registro de cada aluno.

4.3.1 Total de Registros por Curso

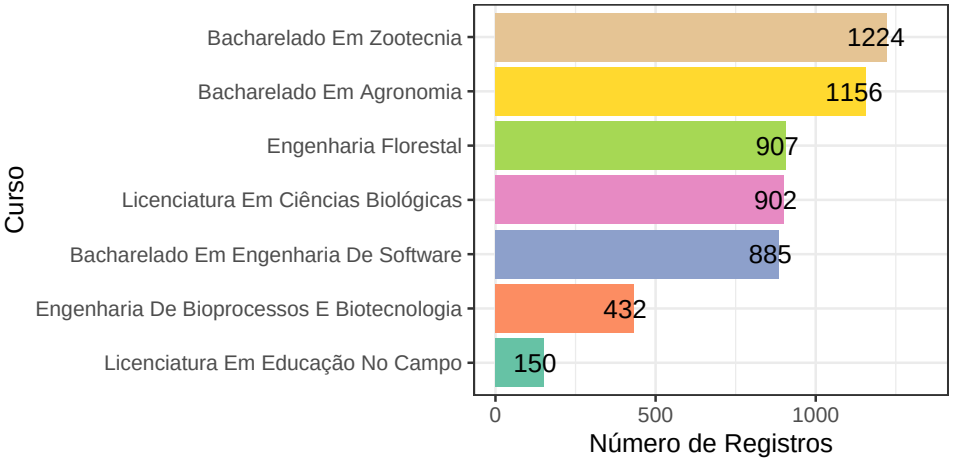
A primeira análise diz respeito à contagem total de registros de alunos no conjunto de dados, segmentada por curso. A Figura 12a ilustra a distribuição total de registros no conjunto de dados completo. Neste cenário, os cursos apresentam um maior número de registros, pois incluem todas as informações acumuladas ao longo do tempo, incluindo transferências, evasões e outros eventos que geraram múltiplos registros para alguns alunos. Por exemplo, no curso de Bacharelado em Zootecnia, há 10013 registros, representando a soma de todas as interações no sistema. Já a Figura 12b reflete a distribuição de registros filtrados, considerando apenas o último registro de cada aluno. Nesse caso, o curso de Bacharelado em Zootecnia, por exemplo,

conta com 1224 registros, indicando o número de alunos únicos que passaram pelo curso ao longo de sua história. A diferença entre os dois números (10013 e 1224) evidencia a quantidade significativa de registros duplicados ou relacionados a mudanças na trajetória acadêmica.

Figura 12 – Comparação do número de registros por curso
(a) Contagem de registros por curso na base de dados completa



(b) Contagem de registros por curso na base de dados que considera apenas os últimos registros



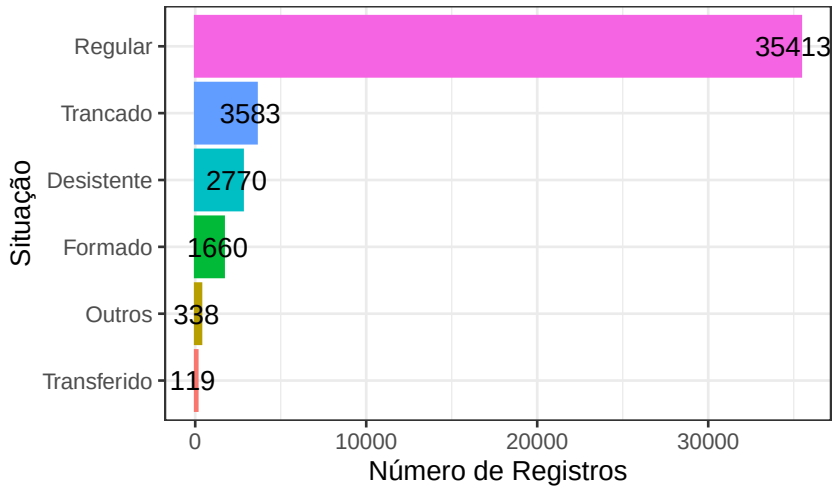
Fonte: Autoria Própria (2024).

4.3.2 Situação dos alunos no sistema acadêmico

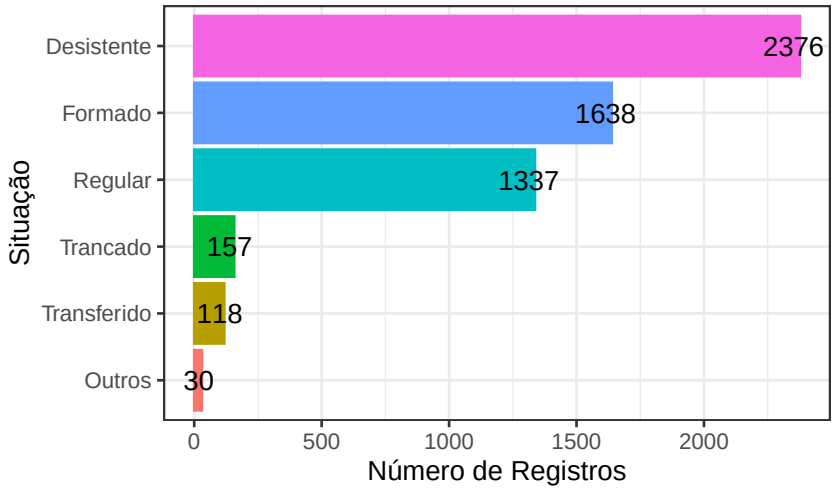
Uma segunda análise refere-se à situação dos alunos no sistema acadêmico. Na Figura 13a, é apresentada a distribuição das situações no *dataset* completo, refletindo todas as ocorrências ao longo do tempo, sem levar em conta a repetição de registros. Já a Figura 13b mostra a contagem das situações considerando exclusivamente o último registro de cada aluno, permitindo observar a distribuição das condições finais dos estudantes, sem redundância de da-

dos anteriores. Essa abordagem mostra novamente a eliminação da redundância de registros intermediários e destaca informações únicas de cada aluno.

Figura 13 – Comparação da contagem de situações
(a) Contagem de situações na base completa



(b) Contagem de situações na base com últimos registros



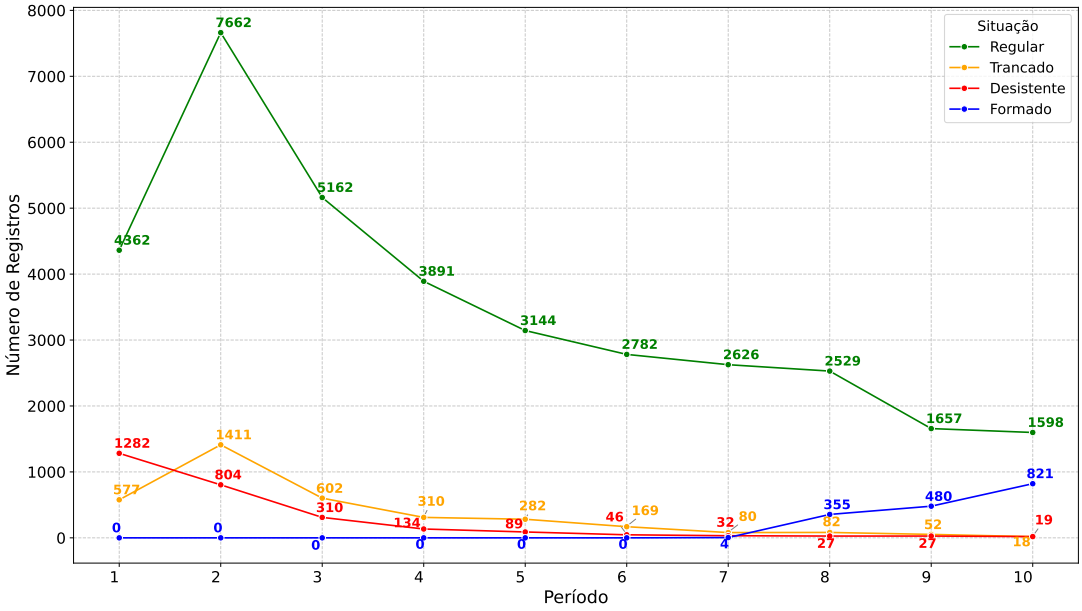
Fonte: Autoria Própria (2024).

Ao olhar para essa contagem, é possível observar uma grande diferença no número de alunos regulares nas duas bases, refletindo o impacto da eliminação de redundâncias no conjunto de dados. Ao analisar a situação "Formado", observa-se que a base completa apresenta 1660 registros, enquanto a base com os últimos registros contabiliza 1638. Essa diferença de 22 registros indica a ocorrência de alunos que concluíram mais de um curso no campus de Dois Vizinhos, evidenciando múltiplas trajetórias acadêmicas bem-sucedidas em diferentes áreas de formação. Observa-se também que o número de alunos rotulados como “Desistente” lidera na base de dados filtrada.

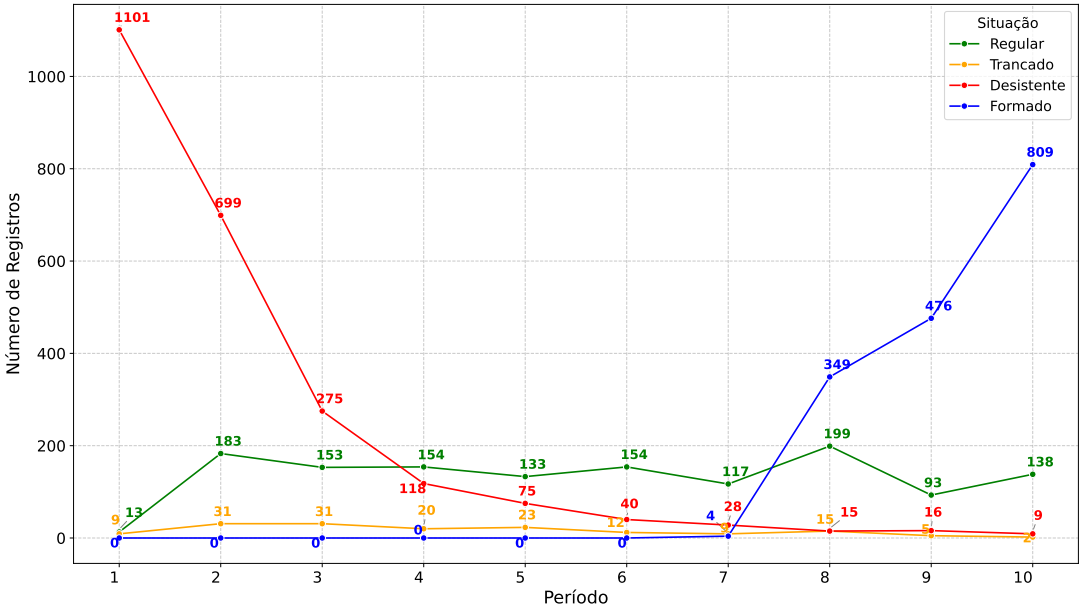
4.3.3 Situação dos alunos por período acadêmico

A análise temporal das situações dos alunos permite identificar tendências importantes ao longo dos períodos acadêmicos. Conforme apresentado na Figura 14a, é possível observar a evolução geral das condições no *dataset* completo, enquanto a Figura 14b fornece uma visão mais específica, focando no último registro de cada aluno. Um aspecto que chama atenção

Figura 14 – Comparação da distribuição das situações por período
(a) Situações por período na base completa



(b) Situações por período na base com últimos registros



Fonte: Autoria Própria (2024).

é a concentração de desistências nos primeiros períodos do curso, indicando desafios iniciais

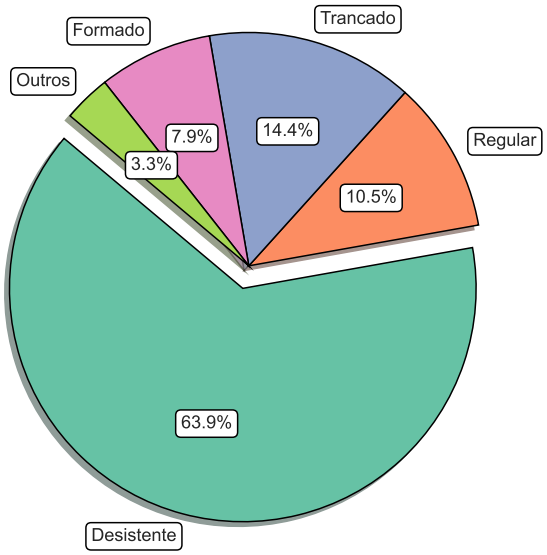
enfrentados pelos estudantes, como dificuldades de adaptação ao ambiente acadêmico, carga de estudos ou questões financeiras. Isso reforça a necessidade de estratégias de acolhimento e suporte mais eficazes nos primeiros semestres, visando aumentar a retenção.

Por outro lado, o número mais elevado de alunos regulares nos períodos iniciais sugere que muitos ingressantes permanecem ativos no início, mas enfrentam barreiras que dificultam sua progressão no curso. Esse padrão pode estar relacionado a fatores como a dificuldade em acompanhar o ritmo acelerado das disciplinas, a complexidade crescente dos conteúdos acadêmicos e os desafios de adaptação às demandas do ensino superior. Muitos estudantes podem encontrar dificuldades em assimilar o conteúdo, especialmente em matérias introdutórias que exigem uma base sólida, além de enfrentar a necessidade de desenvolver habilidades como autonomia nos estudos, gestão do tempo e organização. Essas questões tornam os primeiros semestres críticos para o sucesso acadêmico e reforçam a importância de oferecer suporte pedagógico e estratégias de integração desde o início da trajetória universitária.

4.3.4 Comportamento dos alunos com curso trancado (Trancados)

Quanto aos rótulos, uma análise relevante no conjunto de dados completo diz respeito à situação dos alunos classificados como “Trancado”, atribuído a estudantes que interromperam temporariamente seus estudos, com a intenção de retomá-los em um período subsequente. A Figura 15 apresenta a distribuição percentual das situações atuais desses alunos, que em algum momento da sua trajetória acadêmica trancaram o curso.

Figura 15 – Situações atuais dos alunos que em algum momento trancaram o curso



Fonte: Autoria Própria (2024).

Uma hipótese que pode ser observada é que alunos com o status de “Trancado” apresentam uma maior probabilidade de desistência do curso, por conta da taxa reduzida de re-

torno como alunos regulares. Esse comportamento pode ser interpretado como um indicativo de evasão acadêmica. Assim, o monitoramento dos alunos em situação de “Trancado” se torna uma prática importante, uma vez que permite a identificação precoce de possíveis riscos de evasão, possibilitando a adoção de medidas preventivas para apoiar esses estudantes e, potencialmente, reverter essa situação.

Como parte da análise exploratória, observa-se que a primeira tarefa classificatória, descrita na seção 4.1, apresenta a maior coerência. Nesta tarefa, ao comparar a contagem das situações “Regular + Formado” versus “Desistente + Trancado” para ambos os conjuntos de dados (completo e com os últimos registros), é possível identificar um forte indicativo de que a evasão pode ser prevista com boa precisão até, no máximo, o quinto período. As Figuras 16a e 16b apresentam uma comparação entre a contagem de alunos classificados como regulares e evadidos nesta tarefa, evidenciando um maior desbalanceamento nos últimos períodos.

4.4 Pré-processamento

No pré-processamento dos dados, além de remover todos os identificadores e informações sensíveis, como nome, e-mail e seu registro acadêmico, foram excluídas também atributos constantes para todos os alunos, como: “Campus”, “Sede” e “Funcionamento”, por não contribuírem com informações relevantes para a análise. Em seguida, colunas com mais de 50% de valores ausentes foram descartadas, pois sua inclusão traria pouca ou nenhuma utilidade ao modelo, além de potencialmente introduzir ruído e distorcer os resultados. Por exemplo, as colunas referentes às notas do vestibular foram removidas, pois essas informações se mostraram insuficientes: apenas uma pequena parcela dos alunos presentes na base de dados ingressou por meio de vestibular, um método de admissão relativamente recente na instituição.

Tabela 4 – Colunas removidas por mais de 50% de valores ausentes

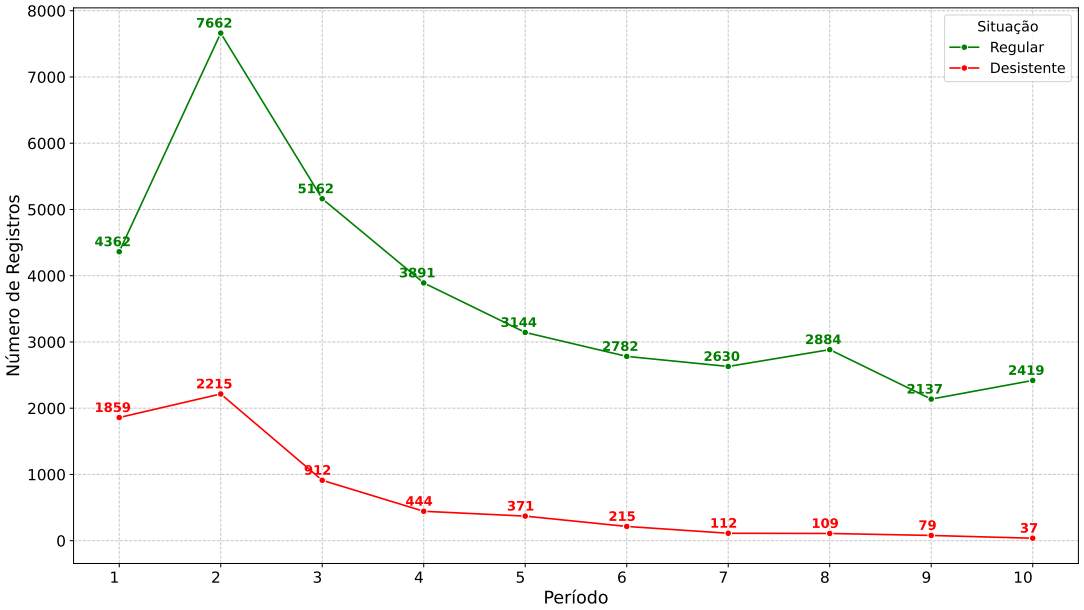
Colunas Removidas	
Categoria Stricto Sensu	Data de Colação de Grau
Nota Vestibular Biologia	Nota Vestibular Filosofia e Sociologia
Nota Vestibular Física	Nota Vestibular Geografia
Nota Vestibular História	Nota Vestibular Literatura Brasileira
Nota Vestibular Língua Estrangeira Moderna	Nota Vestibular Língua Portuguesa
Nota Vestibular Matemática	Nota Vestibular Química

Fonte: Autoria Própria (2024).

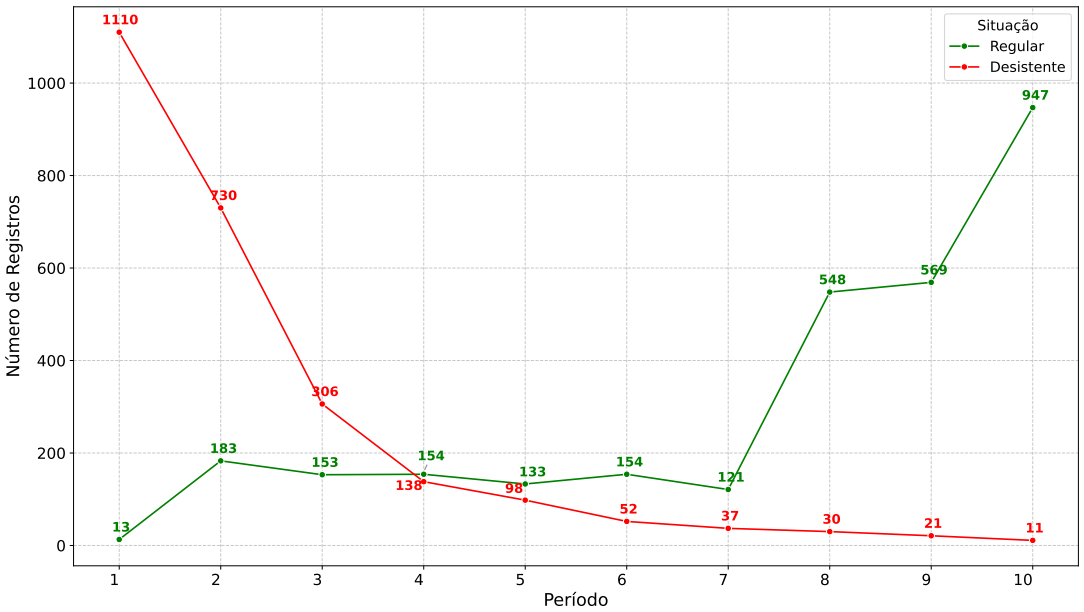
Para as colunas contendo dados categóricos (ou seja, aquelas que incluem texto ou rótulos), foi realizada a conversão para uma representação numérica. Esse procedimento tem como objetivo transformar as categorias em códigos numéricos, facilitando a aplicação de algoritmos de AM implementados na biblioteca *scikit-learn*, que é a referência para AM em *Python*. Embora algoritmos como *DT* e *RF*, que se baseiam em estruturas de árvore, não necessitem dessa conversão, pois conseguem lidar diretamente com variáveis categóricas ao realizar divi-

Figura 16 – Comparação de Situações: Regular + Formado vs Desistente + Trancado

(a) Base de dados com todos os registros



(b) Base de dados apenas com os últimos registros



Fonte: Autoria Própria (2024).

sões com base nos valores dos atributos, tornando essa transformação irrelevante para eles, algoritmos como *NB* e *KNN* exigem a conversão das variáveis categóricas para uma representação numérica, pois dependem de operações que necessitam de variáveis numéricas para calcular distâncias ou probabilidades.

Adicionalmente, vírgulas foram substituídas por pontos, de modo a assegurar o uso do ponto como separador decimal padrão. Caso ocorra falha na conversão de algum valor (por

exemplo, em dados não numéricos), o valor é transformado em *'NaN'* (não numérico), permitindo uma gestão apropriada dos dados ausentes.

Posteriormente, eliminaram-se as colunas quase constantes, aplicando-se um limiar para remoção que permite filtrar atributos sem variação significativa, considerados irrelevantes para o aprendizado dos modelos. Foram testados limiares de constância de $0{,}05$, $0{,}1$, $0{,}15$ e $0{,}2$. Este limiar define o valor mínimo abaixo do qual as colunas são consideradas "quase constantes" e, portanto, removidas. Ou seja, quando a proporção do valor mais frequente em uma coluna ultrapassa o limiar estabelecido, isso indica que os dados nessa coluna não possuem diversidade suficiente para contribuir significativamente para o modelo, resultando na exclusão dessa coluna. A Tabela 5 apresenta os atributos descritivos (colunas) que foram removidos para cada limiar estabelecido.

Tabela 5 – Colunas removidas para cada limiar de constância

Atributo	**Coeficientes de remoção**			
	0.05	**0.1**	**0.15**	**0.2**
Nível de ensino	●	●	●	●
Periodicidade	●	●	●	●
Calouro	●	●	●	●
Coeficiente de rendimento	●	●	●	●
Mudou de curso - mesmo campus	●	●	●	●
Mudou de curso - outro campus	●	●	●	●
Número de entradas no curso	●	●	●	●
Ordem de ingresso no curso	●	●	●	●
Provável jubilamento	●	●	●	●
Regime de ensino	●	●	●	●
Número de entradas em outros cursos		●	●	●
Retenção total			●	●

Fonte: Autoria Própria (2024).

As colunas remanescentes que apresentaram baixa ocorrência de valores ausentes, tiveram seus valores preenchidos usando o tipo da coluna como referência para o método de preenchimento (imputação) (Klosterman, 2020). Atributos numéricos foram preenchidos com a mediana da coluna, enquanto para os dados categóricos foi imputada uma nova categoria específica. A Tabela 6 apresenta as colunas que possuíam valores ausentes, juntamente com o percentual de dados faltantes. Observa-se que cerca de 13% dos alunos da base de dados ingressaram por meio do vestibular próprio da instituição, o que resultou na ausência de informações nas colunas relacionadas às notas do Exame Nacional do Ensino Médio (ENEM). Após a imputação, foi realizada uma nova verificação para identificar e remover variáveis quase constantes. Contudo, nenhuma das colunas que passaram pela imputação apresentou constância abaixo dos limiares utilizados nesse estudo. Dessa forma, todas as colunas imputadas foram mantidas no conjunto de dados para a análise subsequente.

Na etapa final do pré-processamento, foram excluídos atributos altamente correlacionados, utilizando-se um coeficiente de correlação que variou entre $0{,}8$, $0{,}85$, $0{,}9$ e $0{,}95$. Essa

Tabela 6 – Percentual de Valores Ausentes Antes da Imputação

Atributo	Percentual de Valores Ausentes
Nota ENEM Humanas	13,8%
Nota ENEM Natureza	13,8%
Nota ENEM Redação	13,8%
Nota ENEM Linguagem	13,5%
Nota ENEM Matemática	13,5%
Nota final (SiSU/Vestibular)	13,0%

Fonte: Autoria Própria (2024).

abordagem garante que, ao menos, uma coluna de cada par altamente correlacionado seja removida, evitando a duplicação de informações nos modelos. Embora não se escolha uma coluna específica dentro de cada par de colunas correlacionadas, a estratégia visa remover o mínimo necessário para quebrar essas correlações. Em alguns casos, uma coluna pode estar correlacionada com várias outras, formando grupos de colunas inter-relacionadas. Nesse contexto, existem abordagens alternativas que poderiam ser utilizadas, como, por exemplo, a remoção de colunas com menor variabilidade ou a priorização da exclusão de colunas que não são variáveis-chave para o modelo ou para a análise em questão. A Figura 17 ilustra a correlação entre todas as colunas presentes no conjunto de dados anonimizado, já com a remoção das colunas que continham mais de 50% de valores ausentes, enquanto a Tabela 7 apresenta as colunas removidas para cada limiar de correlação analisado.

Tabela 7 – Colunas Removidas para Cada Limiar de Correlação

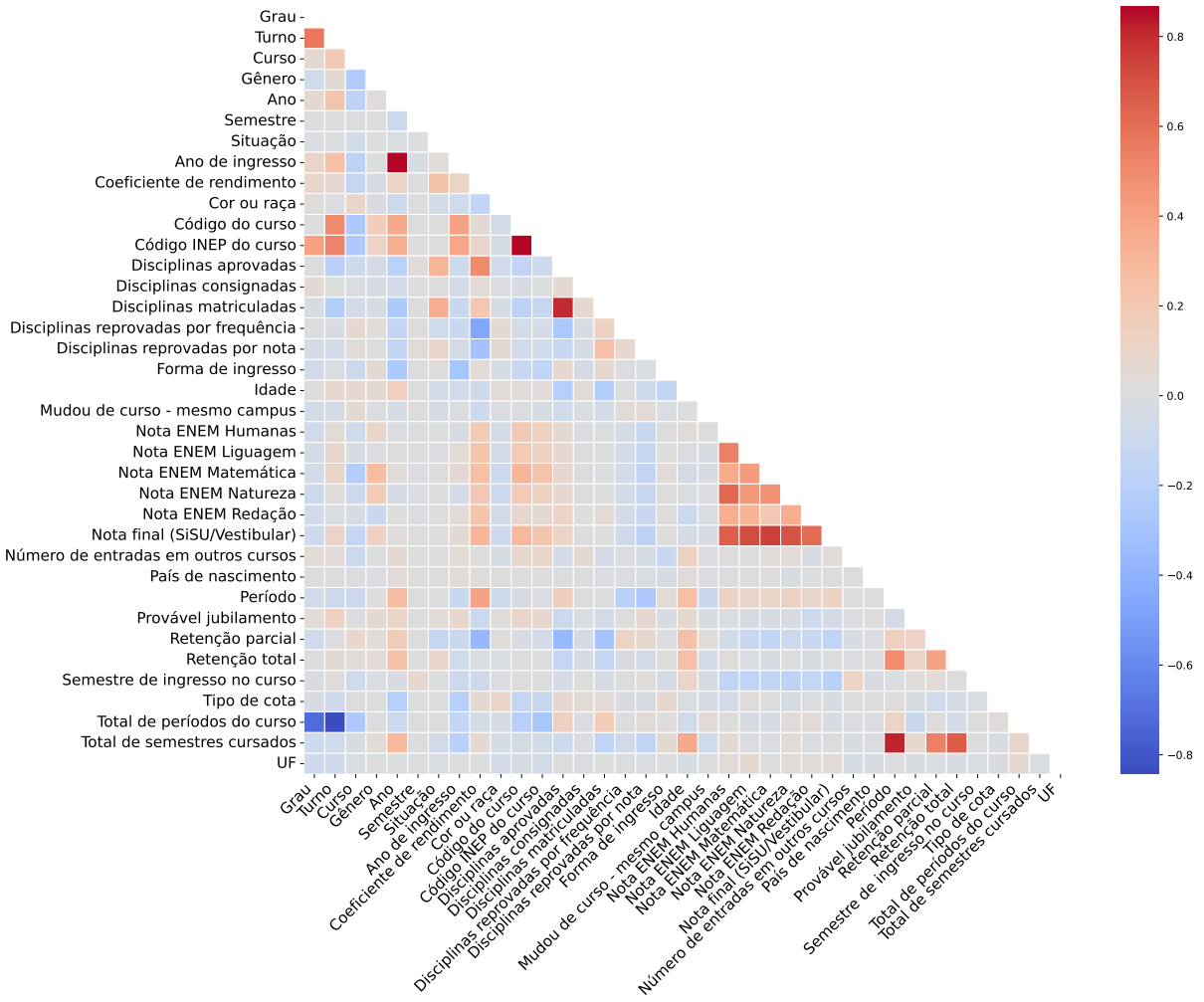
Atributo	Coeficientes de Correlação			
	0.8	0.85	0.9	0.95
E-mail	●			
Total de semestres cursados	●			
Data de nascimento	●			
Código	●	●		
Data de ingresso	●	●	●	●
Código INEP do curso	●	●		
Total de períodos do curso	●			
Ano de ingresso	●	●	●	●
Disciplinas matriculadas	●			

Fonte: Autoria Própria (2024).

4.5 Treinamento dos Modelos Preditivos

A seguir, serão detalhadas as etapas do *setup* experimental adotado para o treinamento e análise dos modelos preditivos. Serão descritas a divisão em conjuntos de treino, validação e teste, bem como os procedimentos de validação cruzada. Além disso, apresenta-se a configu-

Figura 17 – Matriz de Correlação das Características



Fonte: Autoria Própria (2024).

ração do ambiente computacional, os algoritmos de AM selecionados, as métricas de avaliação empregadas e as técnicas estatísticas utilizadas para interpretar os resultados.

4.5.1 Divisão do Conjunto de Dados

Para garantir a reprodutibilidade e a consistência nos experimentos, a divisão de cada conjunto de dados entre treino e teste foi realizada utilizando a validação cruzada estratificada (Castro; Ferrari, 2016). Ela preserva a distribuição original das classes em cada subconjunto gerado, permitindo lidar com possíveis desbalanceamentos presentes nos dados, além de ser útil em cenários onde a quantidade de amostras é limitada, assegurando que a performance do modelo seja avaliada de forma representativa.

Cada experimento foi repetido dez vezes, utilizando diferentes sementes[2] para a inicialização dos processos de embaralhamento e divisão. As sementes utilizadas foram: 145, 278, 392, 49, 203, 411, 89, 356, 27 e 489. O número de partições na validação cruzada (*folds*) foi determinado pelo menor valor entre o total de amostras disponíveis e o limite de dez *folds*. Para conjuntos de dados com menos de duas amostras, a divisão não foi realizada, respeitando as limitações inerentes a esse tipo de análise. A implementação utilizou a classe *StratifiedKFold* da biblioteca *Scikit-learn* configurada com o número de divisões, o embaralhamento dos dados e a semente correspondente.

4.5.2 Algoritmos de AM

Na fase de indução dos modelos preditivos, a seleção dos algoritmos foi orientada pelo critério de interpretabilidade, privilegiando modelos de "caixa branca". Essa escolha visa facilitar a compreensão dos resultados e permitir uma análise mais transparente do comportamento dos modelos. Foram utilizados algoritmos amplamente consagrados na literatura, com base teórica sólida e reconhecida eficiência em diversas aplicações. Os algoritmos empregados nos experimentos incluem: *KNN*, *NB*, *DT* e *RF*. A escolha desses métodos deve-se ao fato de serem os principais algoritmos de caixa branca disponíveis e amplamente utilizados, como evidenciado na revisão bibliográfica. Para garantir uniformidade na execução dos experimentos e considerando as limitações de tempo e recursos computacionais, foram adotados os hiperparâmetros padrão fornecidos pela biblioteca *scikit-learn*, conforme apresentados na Tabela 8.

Tabela 8 – Configuração dos modelos utilizados com a biblioteca *scikit-learn*

Modelo	**Configuração**
K-Nearest Neighbors (KNN)	Número de vizinhos: 5; Métrica de distância: Euclidiana.
Naive Bayes (NB)	Distribuição: Gaussiana.
Decision Tree (DT)	Critério de divisão: Impureza de Gini; Sem limite para profundidade máxima.
Random Forest (RF)	100 árvores no conjunto; Critério de divisão: Impureza de Gini.

Fonte: Autoria Própria (2024).

4.5.3 Métricas de Avaliação

Para avaliar o desempenho dos modelos preditivos, foram utilizadas as seguintes métricas: Acurácia simples, *BAC*, *Precision*, *Recall* e o *F1-Score*. Elas são amplamente empregadas em problemas de classificação, e fornecem uma análise abrangente da performance dos al-

[2] A semente (do inglês, *seed*) é um valor inicial usado para gerar números aleatórios de forma determinística em algoritmos de embaralhamento ou amostragem. Na validação cruzada, ela controla o processo de embaralhamento dos dados antes da divisão em subconjuntos, garantindo que as partições geradas sejam sempre as mesmas em diferentes execuções do experimento.

goritmos. A métrica *BAC* é especialmente útil para lidar com desbalanceamento, pois calcula a média das taxas de verdadeiros positivos de cada classe. A *Precision* e o *Recall* são indicadores complementares que avaliam, respectivamente, a proporção de predições corretas em relação ao total de predições e a capacidade do modelo de identificar corretamente os casos positivos. O *F1-Score*, por sua vez, é a média harmônica entre Precisão e *Recall*, sendo recomendado quando há um compromisso necessário entre ambas as métricas.

Após a aplicação das métricas nos conjuntos de teste, as combinações mais promissoras de *datasets* e configurações de *thresholds* para variáveis correlacionadas e constantes serão avaliadas. A média dos valores obtidos a partir das diferentes sementes será utilizada como indicador de desempenho. Essa abordagem busca identificar quais combinações oferecem melhores resultados no contexto da análise da evasão estudantil, não apenas sob o ponto de vista técnico, mas também considerando sua aplicabilidade gerencial.

Para avaliar o desempenho dos algoritmos, foi utilizado o teste não paramétrico[3] de Friedman. Este teste foi escolhido por não assumir a normalidade dos dados, utilizando as classificações médias dos algoritmos em todas as bases de dados analisadas. A aplicação foi realizada considerando o nível de significância de 95% ($\alpha = 0.05$). Em caso de rejeição da hipótese nula, o teste post-hoc de Nemenyi foi empregado para controlar o erro estatístico e identificar diferenças significativas entre pares de algoritmos, permitindo determinar combinações com desempenho significativamente superior (Barreto *et al.*, 2019). Adicionalmente, o teste de não paramétrico de Wilcoxon foi empregado para comparações entre pares de algoritmos, útil quando se deseja verificar diferenças entre duas condições ou algoritmos, sem pressupor a normalidade dos dados. Na análise final, foi utilizado o teste pareado de Wilcoxon, que considera as diferenças entre duas condições em um mesmo conjunto de dados, ou o teste paralelo, que compara grupos independentes. Todas as análises foram conduzidas considerando um nível de significância de 95% ($\alpha = 0{,}05$).

4.5.4 Paralelismo e Configuração do Servidor

Cada *dataset* será avaliado utilizando todos os algoritmos de AM, combinados com todos os valores de limiares para atributos correlacionados e constantes anteriormente apresentados, além das 10 diferentes sementes mencionadas. Esse processo é conduzido utilizando a técnica de *grid search*[4], que sistematicamente gera todas as combinações possíveis de parâmetros, garantindo uma exploração abrangente do espaço de configurações.

[3] Testes estatísticos não-paramétricos são métodos de análise que não assumem uma distribuição específica para os dados, como a normalidade. Eles são úteis em situações em que os pressupostos dos testes paramétricos (como homogeneidade de variâncias ou linearidade) não são atendidos, ou quando os dados são ordinais ou categóricos.

[4] Grid search é uma técnica sistemática de otimização que avalia todas as combinações possíveis de um conjunto definido de hiperparâmetros, permitindo identificar a configuração que oferece o melhor desempenho em um determinado problema.

O desenvolvimento do *setup* experimental para este projeto foi realizado em duas Interfaces de Desenvolvimento Integrado (IDEs) distintas: *Jupyter Notebook* e *Visual Studio Code (VSCode)*. O *Jupyter Notebook* foi empregado para experimentação e testes rápidos, devido à sua natureza interativa e facilidade de uso. Em contraste, o *VSCode* foi utilizado na criação de *scripts* mais robustos, aproveitando sua ampla gama de funcionalidades e suporte eficaz para o desenvolvimento em *Python*. Os experimentos foram executados em um servidor *HPE ProLiant DL360 Gen10*, equipado com 24 CPUs *Intel Xeon Silver 4214* a 2.20GHz. Para a execução dos algoritmos, foi configurada uma Máquina Virtual, do inglês *Virtual Machine* (*VM*) com 8 núcleos de processamento, 64 GB de RAM, 96 GB de armazenamento e sistema operacional *Ubuntu Server 22.04 LTS*.

Com o objetivo de reduzir o tempo de processamento, adotou-se uma estratégia de paralelismo na *VM*, onde os múltiplos núcleos possibilitaram a execução simultânea de diversas instâncias dos algoritmos, maximizando o aproveitamento dos recursos dos múltiplos núcleos de processamento. A biblioteca *Multiprocessing* foi empregada para distribuir as tarefas de treinamento e teste dos modelos, acelerando o treinamento e reduzindo o tempo total de execução. A implementação prática do paralelismo envolveu a criação de processos independentes para cada execução de treinamento, utilizando filas para a coleta de resultados e a sincronização entre os processos.

A Tabela 9 resume as ferramentas e bibliotecas utilizadas, destacando o objetivo de cada uma. A linguagem de programação adotada foi *Python*, devido à sua versatilidade, vasta gama de bibliotecas e popularidade na comunidade de Ciência de Dados, oferecendo um ambiente robusto e eficiente para a implementação de técnicas de coleta, pré-processamento e análise de dados, além de possibilitar a integração das diversas etapas do processo, desde a manipulação de dados até a aplicação de algoritmos de AM.

4.6 Repositório do Projeto

Com o objetivo de garantir a acessibilidade e fomentar a reprodutibilidade das análises conduzidas neste estudo, foi criado um repositório público no *GitHub*. Nele, estão organizados todos os *scripts*, e *notebooks Jupyter* desenvolvidos ao longo da pesquisa, disponíveis no seguinte endereço: https://github.com/pedromipa/student-dropout-prediction-thesis. Essa iniciativa promove a transparência e a integridade científica, além de oferecer uma base para que outros pesquisadores e interessados possam verificar, reproduzir ou até mesmo aprimorar os métodos e resultados obtidos.

Tabela 9 – Ferramentas e bibliotecas utilizadas na pesquisa

Categoria	**Ferramenta**	**Objetivo**
Linguagem de Programação	*Python*	Implementar os métodos para aquisição, preparação e avaliação dos dados e resultados
Ambiente de Desenvolvimento	*Jupyter Notebook* e *Visual Studio Code*	Escrever, executar e documentar os códigos
Coleta de Dados	Sistema Acadêmico da UTFPR	Obter o conjunto de dados para realizar a pesquisa
Biblioteca para manipular os dados	Pandas	Manipular, limpar, transformar e organizar os dados
Biblioteca de AM	*Scikit-learn*	Dividir o conjunto de dados, implementar e avaliar os algoritmos de AM
Bibliotecas para automatizar processos	*Itertools* e *Multiprocessing*	Automatizar o processo de treinamento, teste e avaliação dos modelos
Biblioteca para Gráficos	*Matplotlib*	Visualizar os resultados através de gráficos e figuras

Fonte: Autoria Própria (2024).

5 RESULTADOS E DISCUSSÕES

Este capítulo apresenta a discussão dos resultados experimentais obtidos. Durante as execuções, foram gerados 115.200 processos individuais, considerando: quatro (4) algoritmos de AM; dez (10) valores distintos de *seeds*, quatro (4) limiares para remoção de atributos constantes, quatro (4) limiares para remoção de atributos correlacionados, e cento e oitenta (180) variações dos conjuntos de dados. Logo: $4 \times 10 \times 4 \times 4 \times 180 = 115.200$ processos executados. Para cada processo foram geradas as médias e desvios padrão das métricas de avaliação extraídas. As próximas seções discutem esses resultados, começando pelas implicações mais gerais, e depois analisando casos mais específicos.

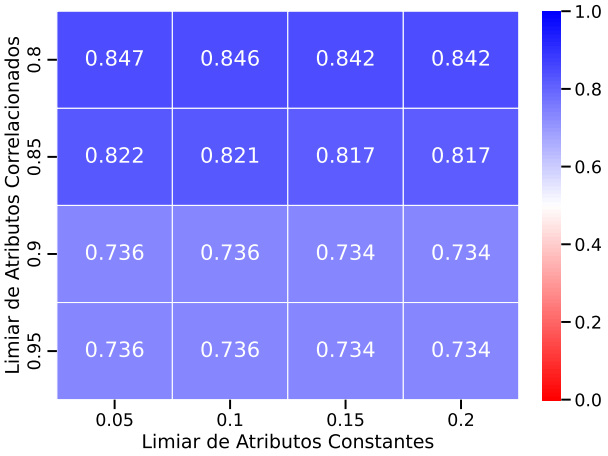
Os resultados são analisados de maneira abrangente ao longo deste capítulo, com o objetivo de definir os melhores limiares para a remoção de atributos constantes e correlacionados, identificar a melhor tarefa classificatória e escolher o algoritmo de AM mais adequado e avaliar as predições realizadas pelos modelos. Cada uma dessas escolhas é discutida com base em critérios de desempenho e impacto nos resultados finais, permitindo uma compreensão das condições que favorecem o melhor desempenho dos modelos. Além disso, são analisadas as predições realizadas por período, comparando as diferentes abordagens.

5.1 Definindo os melhores limiares para remoção de atributos constantes e correlacionados

A primeira análise realizada teve como objetivo determinar os melhores valores de limiares para a remoção de atributos constantes e correlacionados. Essa etapa garante que os dados utilizados nas análises subsequentes sejam relevantes, reduzindo redundâncias e eliminando atributos que pouco contribuem para os modelos. A Figura 18 apresenta a média dos valores de *F1-Score* para cada combinação de limiares aplicados aos experimentos. O eixo X representa os limiares para remoção de atributos constantes, enquanto o eixo Y indica os limiares para remoção de atributos correlacionados. Os valores de *F1-Score* são apresentados exclusivamente para os conjuntos de dados completos, que incluem tanto os registros originais quanto as versões filtradas, contendo apenas os dados mais recentes de cada aluno. Variações adicionais, geradas para os mesmos registros, foram desconsideradas nessa análise. As áreas mais azuis representam combinações de limiares que geraram os melhores desempenhos, enquanto as áreas vermelhas indicam as combinações que resultaram em um desempenho inferior.

Os resultados apresentados na figura evidenciam que o impacto do limiar aplicado à remoção dos atributos constantes é limitado, mostrando variações mínimas nos valores de *F1-Score*. Em contrapartida, os limiares aplicados para remoção de atributos correlacionados mostraram-se significativamente mais influentes, com diferenças substanciais no desempenho dependendo do valor escolhido. Por exemplo, ao ajustar o limiar para valores mais elevados,

Figura 18 – Valores médios de *F1-Score* obtidos nos experimentos com diferentes valores de limiares



Fonte: Autoria Própria (2024).

observou-se uma redução no *F1-Score*, indicando que a remoção de atributos altamente correlacionados impacta consideravelmente a performance do modelo. A combinação dos limiares de $0{,}8$ para correlacionados, e $0{,}05$ para constantes obteve a melhor média de resultados.

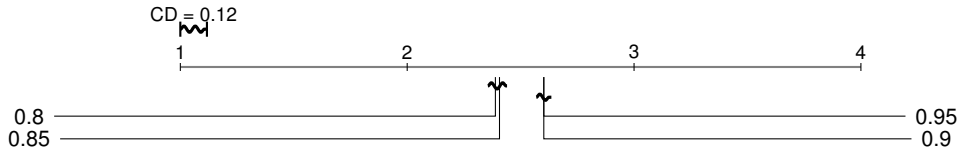
Para identificar diferenças estatísticas, foi aplicado o teste de Friedman-Nemenyi, que avalia quais pares de algoritmos apresentam diferenças estatisticamente significativas. O teste considera que dois limiares são significativamente diferentes quando as médias de seus *ranks* diferem por pelo menos um valor crítico, denominado *Critical Difference* (Diferença Crítica, do inglês *Critical Difference* (*CD*)). No contexto deste estudo, os limiares foram ranqueados do melhor para o pior desempenho, sendo que menores valores de *rank* indicam melhor desempenho. Os resultados são visualmente representados por linhas onduladas que conectam os pares de limiares. Quando dois limiares estão conectados, indica que suas diferenças de desempenho não são estatisticamente significativas. Por outro lado, quando não há conexão, pode-se inferir, com 95% de confiança, que existe uma diferença estatística entre os algoritmos.

Na Figura 19a observa-se que os algoritmos com resultados $0{,}8$ e $0{,}85$ são estatisticamente equivalentes entre si, mas apresentam desempenho significativamente melhor do que os algoritmos com valores $0{,}95$ e $0{,}9$, os quais removem mais atributos. Na Figura 19b, os valores $0{,}1$ e $0{,}05$ também são equivalentes, mas têm desempenho significativamente superior em relação a $0{,}2$ e $0{,}15$.

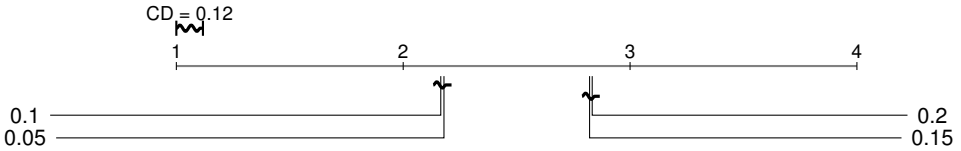
A escolha dos valores $0{,}8$ e $0{,}05$ como melhores configurações foi baseada em uma análise que considerou tanto o desempenho estatístico quanto a consistência dos resultados. Embora o valor $0{,}1$ tenha ficado em primeiro lugar no ranking do teste de Nemenyi, sua média apresentou uma diferença inferior de apenas $0{,}001$ em relação ao valor $0{,}05$ na média de resultados obtidos, conforme a Figura 18. Essa diferença, apesar de destacada no ranking, não é estatisticamente significativa, conforme indicado pelo teste pós-hoc, que apontou equivalência entre $0{,}1$ e $0{,}05$. A escolha de considerar a remoção dos atributos com constância superior a 5% se justifica, portanto, por sua proximidade em desempenho com o limiar de remoção de

Figura 19 – Diagrama de Diferença Crítica (CD) comparando os diferentes limiares utilizados nos experimentos

(a) Comparação dos atributos correlacionados



(b) Comparação dos limiares de atributos constantes



Fonte: Autoria Própria (2024).

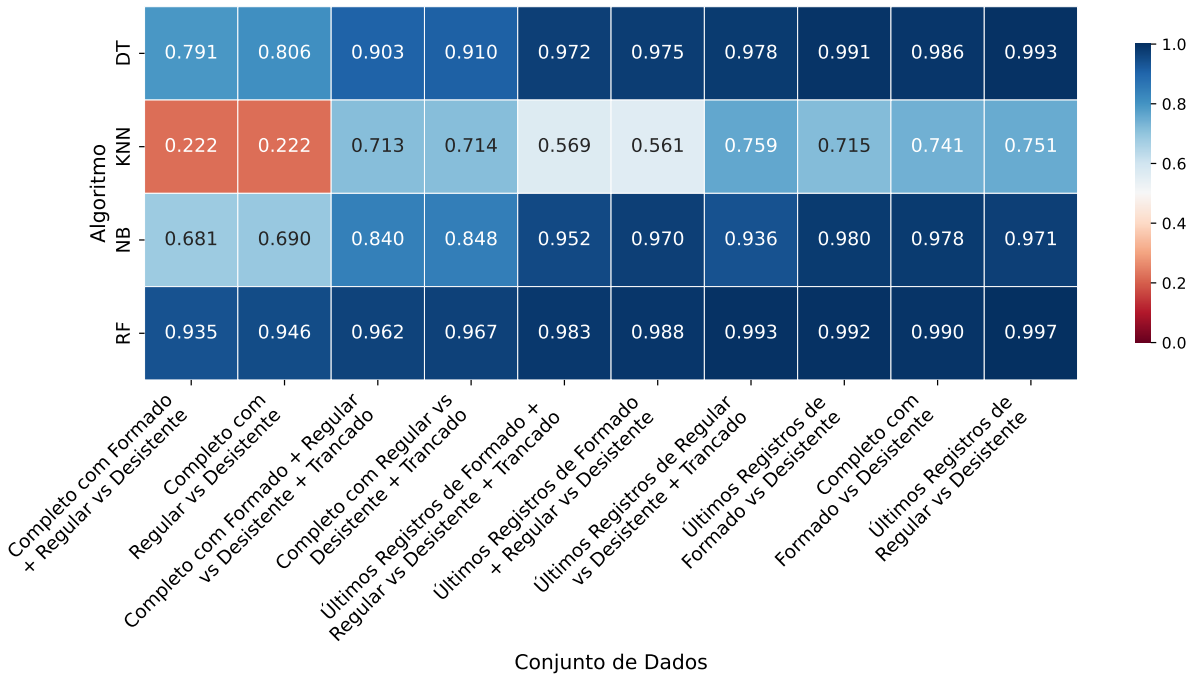
atributos constantes de $0{,}1$, aliada à remoção de um atributo a menos, conforme mostrado na Tabela 5, mantendo a variável *Número de entradas em outros cursos*, que pode oferecer informações relevantes para a análise ao dizer quantas vezes o aluno ingressou em diferentes cursos de graduação no campus de Dois Vizinhos.

5.2 Definindo a melhor tarefa preditiva

Como próxima etapa, buscou-se identificar a tarefa preditiva com melhor desempenho entre os cinco tipos de tarefas definidos na metodologia na seção 4.2 e mostradas na Tabela 3. A Figura 20 exibe um mapa de calor com os resultados obtidos para cada tarefa, tanto no conjunto de dados completo (que inclui todos os registros de cada aluno) quanto no conjunto filtrado (que considera apenas os últimos registros de cada aluno). A análise dos resultados revela padrões distintos entre esses dois conjuntos de dados. O gráfico organiza os valores de *F1-Score* de maneira sistemática, com as tarefas dispostas da esquerda para a direita de acordo com a média dos resultados obtidos, facilitando a comparação detalhada de desempenho e a identificação dos melhores resultados obtidos.

Observa-se, por exemplo, que a tarefa relacionada à classificação "Formado + Regular versus Desistente", utilizando o conjunto de dados completo, apresenta valores atípicos, com *F1-Score* variando entre $0{,}2$ e $0{,}3$, distantes da faixa usual. Para o conjunto contendo apenas os registros mais recentes dos alunos, as tarefas preditivas apresentaram desempenhos relativamente uniformes, com baixa dispersão nos valores de *F1-Score*. Esse comportamento é atribuído ao maior balanceamento de classes nesse conjunto. Por outro lado, no conjunto completo, duas tarefas se destacaram devido à maior variação nos resultados, evidenciando uma sensibilidade acentuada à diversidade dos dados. Essa variação incluiu baixos valores de *F1-Score* obtidos pelos algoritmos *NB* e *KNN*, o que alterou a média obtida. Esse efeito pode ser

Figura 20 – Valores médios de F1-Score obtidos por todos os algoritmos treinados em todas as tarefas preditivas



Fonte: Autoria Própria (2024).

explicado pelo maior desbalanceamento de classes e por um possível viés temporal introduzido ao considerar o conjunto de dados completo.

Os resultados apontam que o melhor desempenho foi alcançado com o conjunto de dados contendo apenas os últimos registros dos alunos, na tarefa classificatória “Regular vs Desistente”. Essa tarefa destacou-se como a mais precisa, apresentando um *F1-Score* médio de 0.997 com o *RF*. No entanto, vale ressaltar que esta tarefa também possui o menor número de exemplos, devido à exclusão de cerca de 15 mil amostras no conjunto completo e mais de mil no conjunto filtrado.

O segundo melhor desempenho foi alcançado com o conjunto de dados completo, na tarefa classificatória “Formado vs Desistente”. Embora essa tarefa também exclua uma grande quantidade de registros disponíveis, ela consegue equilibrar o balanceamento entre as classes, o que contribui para resultados mais consistentes. No conjunto completo, o número de registros da classe “Regular” é desproporcionalmente elevado, uma vez que o mesmo aluno pode aparecer várias vezes como “Regular” em diferentes períodos. Ao remover os alunos classificados como “Regular” e manter apenas os registros das classes “Formado” e “Desistente”, o volume de dados no conjunto completo sofre uma redução superior a 90%, resultando em um número significativamente menor de registros, conforme detalhado na Tabela 3. Essa redução impacta a complexidade da tarefa, tornando-a mais direcionada e controlada, mas, ao mesmo tempo, limita a generalização dos resultados para cenários mais amplos.

Entre as tarefas analisadas, a tarefa 5 (“Regular vs Desistente”) destacou-se como a mais simples, alcançando altos valores de *F1-Score* em todos os algoritmos. No entanto, sob a perspectiva de aplicabilidade prática, a tarefa 1 (“Formado + Regular vs Desistente + Trancado”) revela-se mais adequada para um sistema de apoio à decisão. Essa tarefa utiliza todos os exemplos relevantes disponíveis no Sistema Acadêmico e reflete cenários mais abrangentes relacionados ao abandono escolar. Além disso, aborda situações de maior relevância para a gestão acadêmica, sendo fundamental para monitorar e mitigar problemas associados à evasão. Nesta tarefa, o *RF* também apresentou excelente desempenho, com um *F1-Score* médio de $0{,}983$. Dada a importância e a abrangência da tarefa 1, as análises apresentadas nas próximas seções focarão nesse cenário, utilizando seus resultados como padrão para a discussão e interpretação.

5.3 Definindo o melhor algoritmo de AM

Um passo posterior que pode ser realizado é definir qual algoritmo é mais apropriado para o conjunto de tarefas definido. Analisando o desempenho dos algoritmos avaliados em todas as instâncias de experimentos, observa-se que o *KNN* apresentou o pior valor médio de *F1-Score*, com $0{,}59$, seguido pelo *NB* e pelo *DT*, cujos valores de *F1-Score* foram $0{,}81$ e $0{,}92$, respectivamente. Os resultados obtidos pela *DT* são especialmente notáveis, considerando que este é um modelo relativamente simples. Além de oferecer um desempenho competitivo, ele apresenta como vantagem a capacidade de retornar um conjunto de regras interpretáveis. Essa característica é particularmente útil para os gestores acadêmicos, pois permite identificar as causas da evasão ou situações de iminência de abandono por parte de um aluno.

O melhor desempenho foi alcançado pelo algoritmo *RF*, com um *F1-Score* médio de $0{,}97$, superando consistentemente todos os outros algoritmos nas cinco tarefas. Para avaliar a equivalência entre os algoritmos com base nos valores de *F1-Score*, novamente foi aplicado o teste estatístico não paramétrico de Friedman com $\alpha = 0{,}05$ (95%). A hipótese nula do teste, que postula a inexistência de diferenças significativas no desempenho entre os algoritmos, foi rejeitada, indicando que há diferenças estatisticamente significativas nos desempenhos dos algoritmos analisados. A Figura 21 apresenta o diagrama de *CD* resultante, onde algoritmos conectados indicam ausência de diferenças estatisticamente significativas entre o *RF* e *DT*. Observa-se que o *RF* é consistentemente superior, com diferenças significativas em relação ao *NB* e *KNN*.

Figura 21 – Diagrama de Diferença Crítica (CD) comparando os algoritmos usados nos experimentos

Fonte: Autoria Própria (2024).

O fraco desempenho do *KNN* pode ser atribuído à sua dependência da proximidade entre os pontos de dados, o que torna o algoritmo sensível à densidade e à variabilidade dos dados. Essa característica foi particularmente desafiadora no conjunto completo, que apresenta maior complexidade e possíveis redundâncias. O *NB*, que assume independência entre as variáveis, enfrentou dificuldades ao lidar com relações interdependentes mais complexas, comuns em conjuntos de dados reais. O *DT*, embora mais flexível, demonstrou limitações em cenários com alta dimensionalidade, o que comprometeu seu desempenho nas tarefas de “Formado + Regular vs Desistente” e “Regular vs Desistente” no conjunto de dados completos, algo que pode ser atribuído ao alto desbalanceamento de classes presentes em ambas. Em contraste, o *RF* demonstrou um desempenho estável e confiável em todas as tarefas analisadas. Sua abordagem baseada em *ensemble*, que combina múltiplas árvores de decisão treinadas em subconjuntos aleatórios de dados e variáveis, permitiu lidar eficazmente com a heterogeneidade dos dados. Essa característica robusta torna o RF especialmente adequado para contextos de alta variabilidade e complexidade.

5.4 Análise das predições realizadas por período

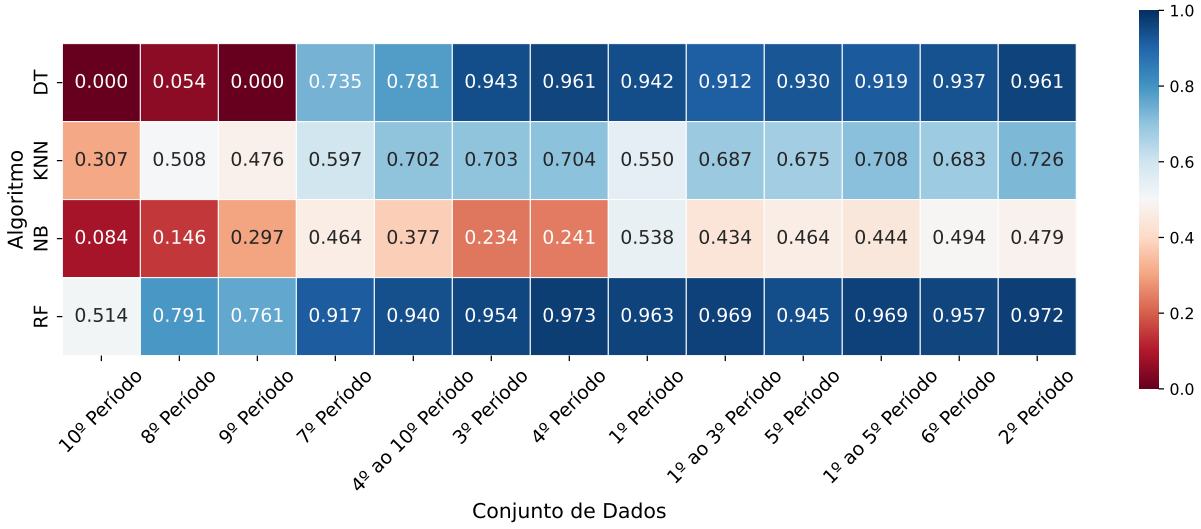
Anteriormente, constatou-se que é possível prever a evasão utilizando as variáveis disponíveis no sistema. Contudo, verificou-se uma tendência relevante: alunos matriculados nos períodos iniciais, especialmente entre o 1º e o 3º, apresentam maior predisposição à evasão. A partir dessa constatação, formulou-se a hipótese de que a decomposição do conjunto de dados em diferentes versões, considerando registros correspondentes a períodos acadêmicos específicos, poderia revelar variações no desempenho dos modelos preditivos e identificar cenários nos quais esses modelos apresentam maior eficácia.

Para avaliar essa hipótese, foram geradas diferentes configurações de conjuntos de dados: inicialmente, conjuntos individuais para cada período acadêmico, abrangendo do 1º ao 10º períodos; em seguida, combinações de períodos específicos, como os períodos iniciais agrupados (do 1º ao 3º e do 1º ao 5º) e os períodos intermediários e finais agrupados (do 4º ao 10º). As figuras apresentadas a seguir ilustram o desempenho médio dos algoritmos em cada configuração analisada.

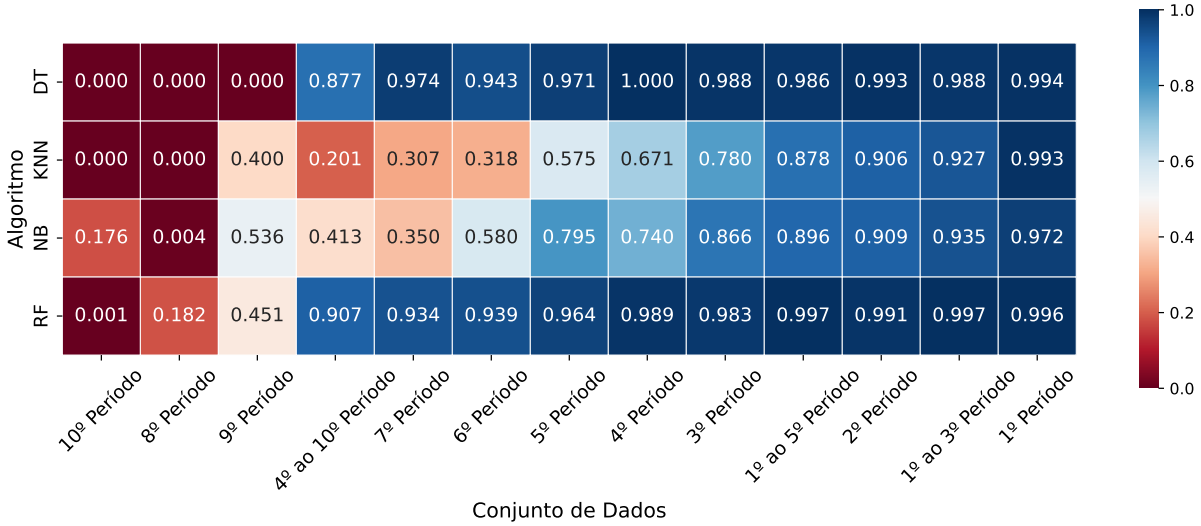
A Figura 22a apresenta um *heatmap* com os valores médios de *F1-Score* obtidos para os dados completos, considerando os diferentes períodos. O eixo x lista os *datasets* de acordo com os valores de *F1-Score* médios obtidos por todos os algoritmos. Os *datasets* estão ordenados do pior para o melhor, da esquerda para a direita. Quanto mais azul, mais próximo de $1{,}0$ são os valores de *F1-Score*, e melhor são os modelos induzidos. De maneira contrária, quanto mais vermelho, mais próximo de zero são as performances, e piores os modelos preditivos.

Já a Figura 22b mostra os resultados para o conjunto de dados filtrados, que considera apenas os últimos registros de cada aluno. Ao analisar a figura, é possível observar o comportamento geral dos algoritmos ao longo dos períodos acadêmicos, identificando os períodos

Figura 22 – Mapas de calor dos valores médios de *F1-Score* por período
(a) Dataset com todos os registros



(b) Dataset com os últimos registros



Fonte: Autoria Própria (2024).

que resultam em maior ou menor desempenho preditivo. De maneira geral, os períodos iniciais, como o 1º e o 2º, tendem a ser mais fáceis de prever devido ao maior número de alunos evadidos nesses estágios. Esse alto número de casos positivos fornece um maior conjunto de amostras para os modelos aprenderem os padrões associados à evasão. Os períodos mais avançados apresentam maior dificuldade de predição devido à menor quantidade de alunos evadidos, o que resulta em conjuntos de dados desbalanceados. Essa característica frequentemente prejudica o aprendizado do modelo, pois há uma insuficiência de amostras representativas. Um exemplo disso é o décimo período, que possui poucos registros de evasão, como ilustrado nas Figuras 16a e 16b.

A comparação entre os resultados para os dados completos e aqueles filtrados pelos últimos registros de cada aluno também revela diferenças importantes. Nos dados filtrados, há um maior balanceamento entre as classes, uma vez que apenas os registros mais recentes de cada aluno são considerados, reduzindo a redundância de informações relacionadas a períodos anteriores. Esse maior equilíbrio entre as classes pode facilitar o aprendizado de alguns modelos, mas, ao mesmo tempo, diminui a riqueza dos dados disponíveis para identificar padrões temporais.

O desempenho consistente do algoritmo *RF* em quase todos os cenários pode ser explicado por algumas características intrínsecas ao método. Primeiro, o *RF* é menos sensível ao desbalanceamento de classes devido à natureza dos seus critérios de divisão nos nós das árvores, que tendem a priorizar a redução da impureza global no conjunto de dados. Além disso, o *RF* é menos afetado pelo viés temporal dos dados, uma vez que suas árvores baseiam-se em amostras do conjunto de treino, criando uma diversidade que mitiga os efeitos de eventuais correlações temporais entre as amostras.

Após a análise inicial com mapas de calor, optou-se em seguir com a utilização de gráficos de linha, devido à sua eficácia em representar tendências temporais e variações no desempenho dos algoritmos ao longo dos períodos analisados, permitindo visualizar, de forma mais clara e direta, flutuações nos resultados, além de facilitar a identificação de padrões específicos e possíveis mudanças no comportamento dos modelos ao longo do tempo.

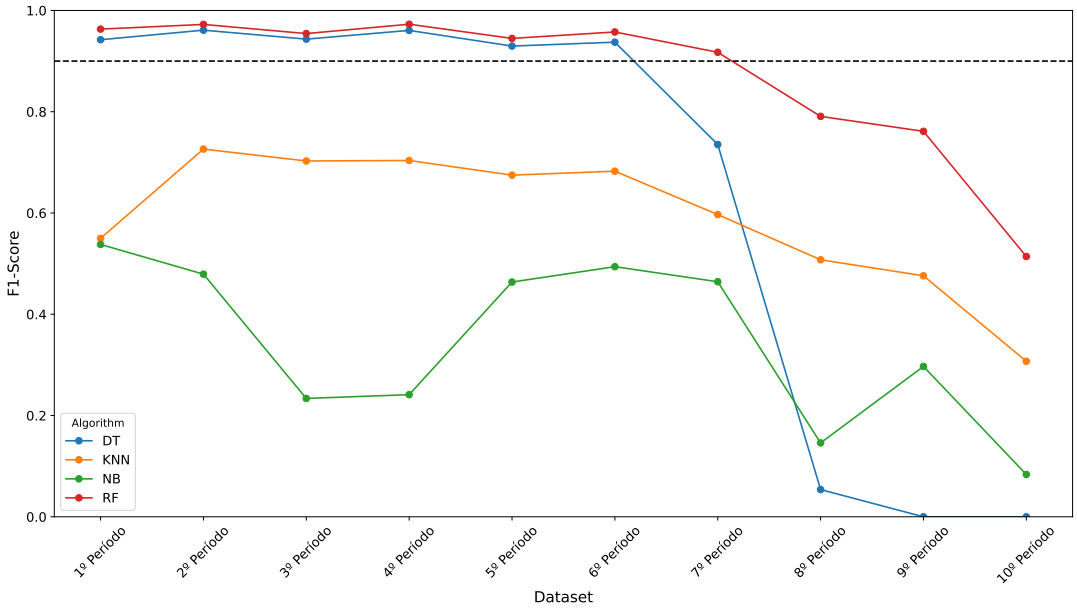
Em termos de desempenho por período específico, a Figura 23a revela uma leve queda no desempenho do *RF* no sétimo período, mas o algoritmo mantém resultados relativamente aceitáveis nos períodos subsequentes, como o oitavo e nono. Esse comportamento indica que, embora o desempenho sofra uma pequena variação ao longo dos períodos, o *RF* mantém uma boa performance geral.

A Figura 23b indica uma queda acentuada no desempenho entre o sétimo e o oitavo período, com uma redução significativa nos valores de *F1-Score*. Além disso, a análise das quedas nos desempenhos do *KNN* e do *NB* à medida que os períodos avançam sugere que esses algoritmos são mais sensíveis à variação nos dados, especialmente em períodos mais distantes.

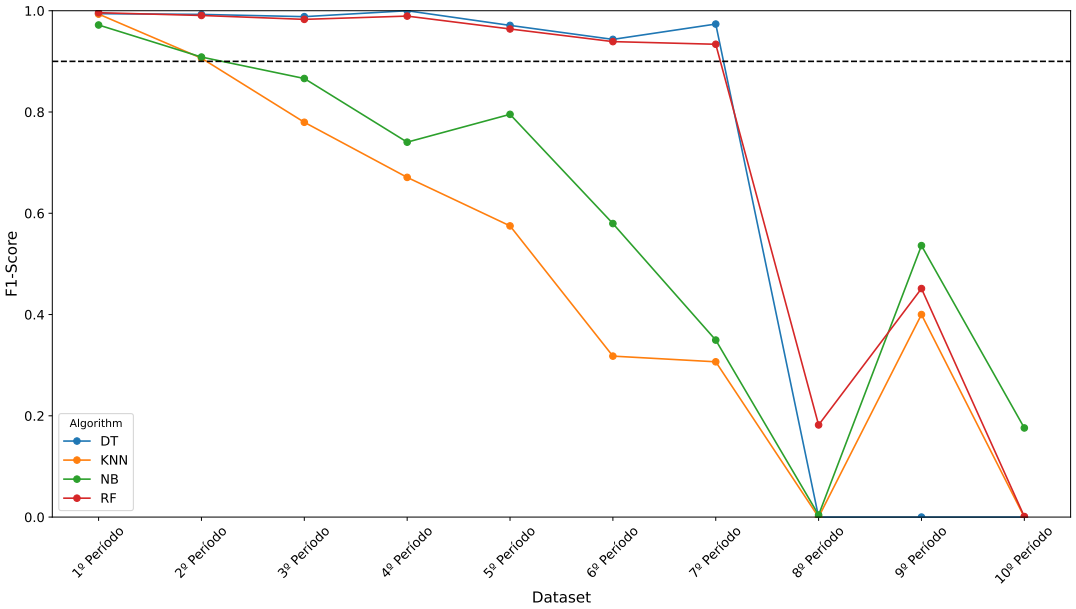
Ao agrupar os períodos, os gráficos das Figuras 24a e 24b reforçam a robustez do *RF*, que mantém um desempenho consistente e elevado em todos os cenários. Essa análise é relevante porque a estabilidade do *RF* em diferentes períodos demonstra sua capacidade de generalização, indicando que ele é menos suscetível a variações nos dados ao longo do tempo. Essa característica é especialmente importante em cenários onde a distribuição dos dados pode mudar de um período para outro, como em aplicações práticas que envolvem séries temporais ou sistemas sujeitos a sazonalidade.

Além disso, o bom desempenho em períodos iniciais e intermediários, como do quarto ao décimo, sugere que o *RF* pode ser uma escolha confiável para aplicações que exigem consistência e precisão em análises preditivas, independentemente do momento em que os dados

Figura 23 – Comparação de valores médios de *F1-Score* obtidos por período
(a) Dataset com todos os registros



(b) Dataset com os últimos registros



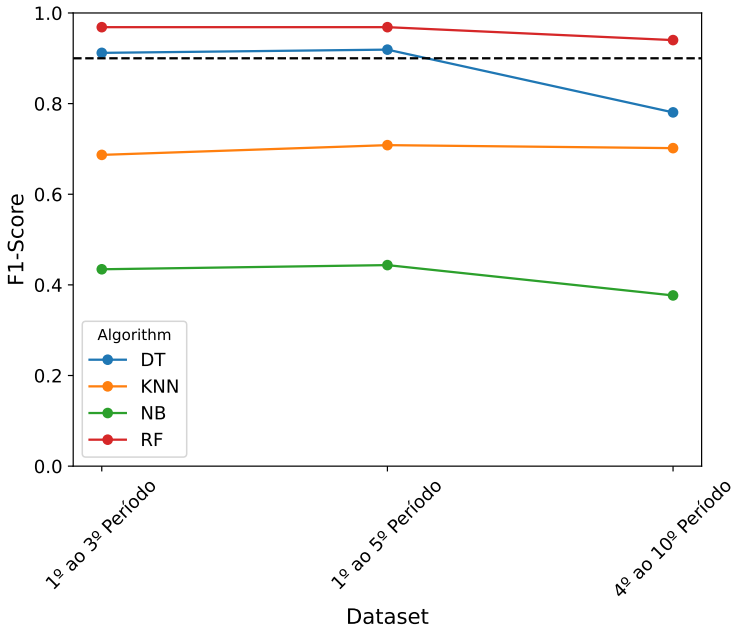
Fonte: Autoria Própria (2024).

foram gerados. Essa estabilidade pode implicar maior confiabilidade nas decisões baseadas no modelo, mesmo em contextos de maior incerteza ou com dados históricos diversos.

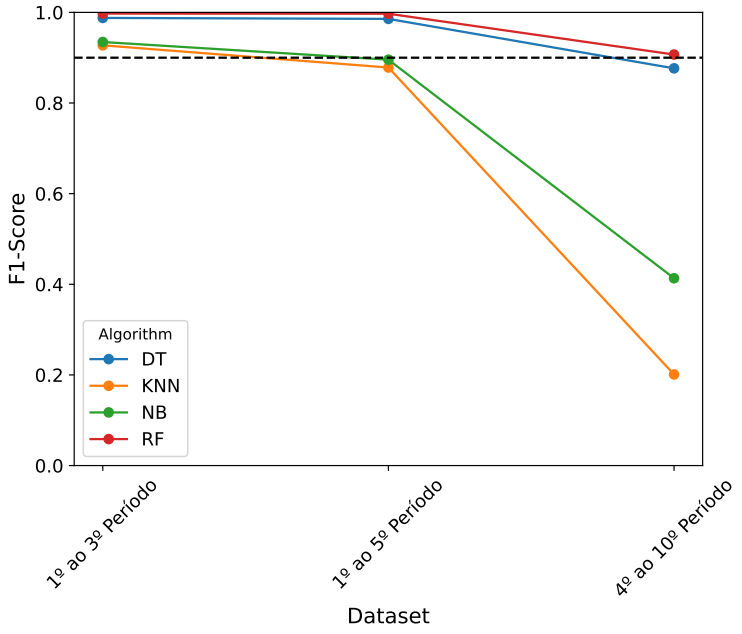
5.5 Predições dos melhores modelos

Nas seções anteriores pudemos analisar experimentalmente, e concluir que:

Figura 24 – Comparação dos resultados obtidos nos períodos agrupados
(a) Dataset com todos os registros



(b) Dataset com os últimos registros



Fonte: Autoria Própria (2024).

- a melhor tarefa preditiva é a tarefa que define como classes as situações dos alunos: “Formado + Regular vs Desistente + Trancado”;
- na análise de limiar de correlação, o melhor valor experimental foi $0{,}8$, que mantem performance e torna a base de dados mais enxuta;

- o mesmo pode-se argumentar do valor de limiar para remoção de atributos constantes igua a $0{,}05$; e
- por fim, o melhor algoritmo preditivo foi o algoritmo *Random Forest* (*RF*).

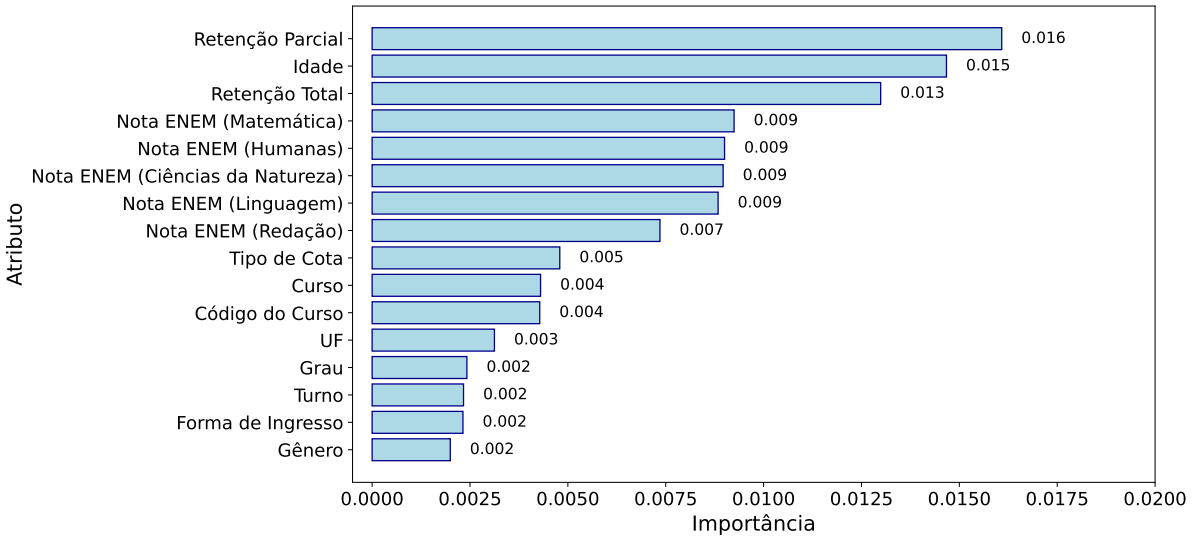
Sendo assim, levando em consideração esse setup experimental, analisamos as predições obtidas pelo melhor algoritmo, na melhor tarefa, usando os melhores valores de limiares para pré-processamento da base de dados. Uma primeira análise interessante, e inerente aos modelos de RF, é olhar a importância relativa dos atributos avaliados pelos modelos induzidos. Estes valores, fornecem uma visão detalhada sobre os fatores considerados pelo modelo para a tomada de decisão. A análise de importância das características foi realizada tanto nos dataset com os registros completos quanto na versão com apenas os últimos registros dos alunos. A Figura 25 ilustra as variáveis mais relevantes, cujas importâncias, em termos do índice de Gini, são superiores a $0{,}01$. Os valores foram obtidos a partir da média de 10 execuções de todos os *seed*.

No caso do conjunto de dados com todos os registros, a variável “Retenção Parcial” se destacou como o melhor indicador, seguida de “Idade” e “Retenção Total”. Isso evidencia que alunos retidos em períodos intermediários, possivelmente devido a uma carga horária insuficiente ou dificuldades acadêmicas, têm maior probabilidade de abandonar o curso. Por outro lado, ao analisar o conjunto de dados contendo apenas os últimos registros dos alunos, observou-se que o modelo atribuiu maior importância à variável “Idade” e “Retenção Total”, com “Retenção Parcial” aparecendo apenas na oitava posição. Essa mudança indica que, em estágios mais avançados, a idade do aluno e o histórico acumulado de reprovações tornam-se fatores mais influentes na predição de desistência.

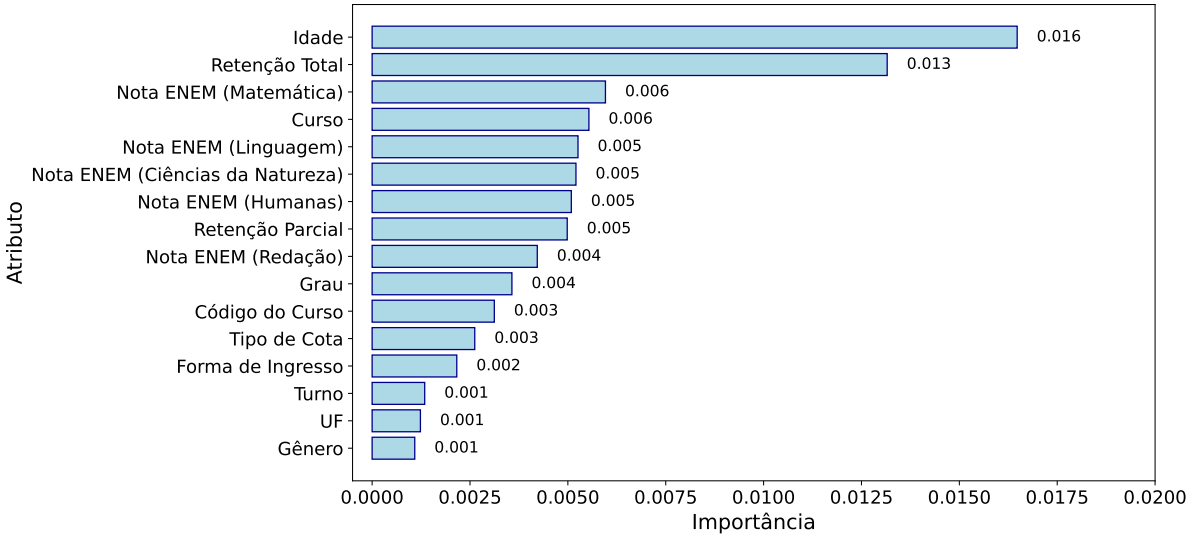
Uma análise mais abrangente mostra semelhanças e diferenças importantes entre os dois rankings de relevância. Em ambos os cenários, “Nota ENEM (Matemática)” aparece consistentemente entre os atributos mais importantes, enquanto outras notas do ENEM, como “Ciências da Natureza”, “Humanas” e “Linguagem”, apresentam relevância intermediária. Esse padrão sugere a possibilidade de inferir as afinidades do estudante com os conceitos centrais do curso, apontando para potenciais conexões entre o perfil acadêmico e as exigências específicas das áreas de estudo. Além disso, variáveis como “Gênero”, “UF” e “Turno” aparecem nas últimas posições em ambos os conjuntos, indicando que têm menor impacto no desempenho do modelo. Isso pode ser explicado pelo fato de que essas variáveis não possuem uma relação direta e significativa com o desempenho acadêmico ou com os fatores que levam à evasão. Diferentemente das notas do ENEM, que refletem habilidades específicas do estudante, essas variáveis têm caráter mais contextual ou demográfico, e seu impacto pode estar diluído ou mediado por outros fatores mais determinantes no modelo. Ao comparar os dois rankings, observa-se que algumas variáveis, como “Curso” e “Código do Curso”, têm posições relativamente consistentes, sugerindo uma influência moderada e constante nos dois conjuntos.

Figura 25 – Importância relativa dos atributos estimados pelos modelos de *RF*

(a) Dataset com todos os registros



(b) Dataset com os últimos registros



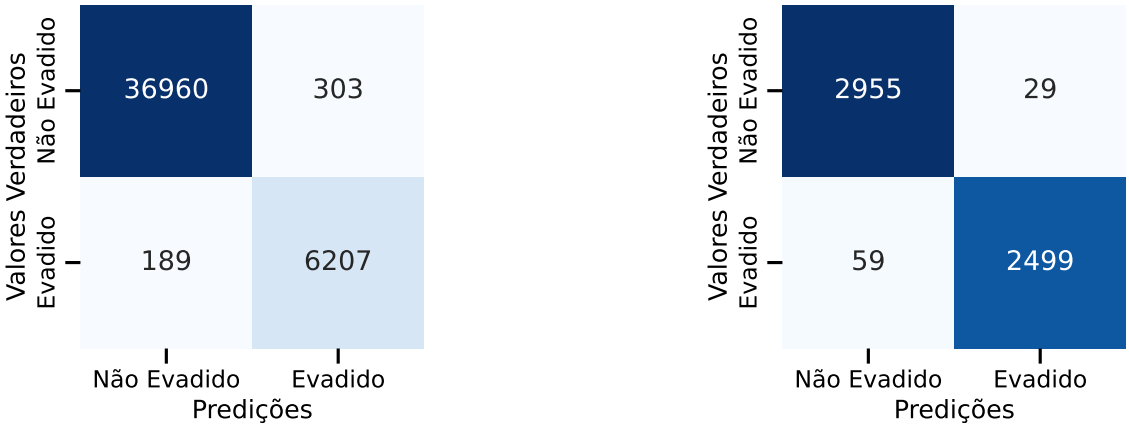
Fonte: Autoria Própria (2024).

Esses resultados indicam que o modelo *RF* consegue identificar padrões distintos dependendo da janela temporal dos registros analisados, refletindo mudanças nos fatores que influenciam a desistência ao longo do tempo acadêmico. A 26 ilustra essas observações, apresentando as matrizes de confusão médias do RF para ambos os conjuntos. Notam-se diferenças claras no desempenho entre os dois cenários, que refletem a variação na importância das variáveis preditivas e destacam a necessidade de adaptar a análise às características específicas de cada etapa do percurso acadêmico.

As matrizes revelam que ambas as classes apresentaram altas taxas de acurácia, com uma *Precision* de $0{,}955$ e $0{,}988$, e uma *BAC* de $0{,}981$ e $0{,}983$, respectivamente. O alto número de exemplos presentes no conjunto de dados contribui significativamente para a definição

Figura 26 – Matrizes de confusão obtidas pelo algoritmo *RF*

(a) Dataset com todos os registros **(b) Dataset com últimos registros**



Fonte: Autoria Própria (2024).

precisa dos limites de decisão do modelo. Neste contexto, a classe positiva, que corresponde aos estudantes “Desistente”, apresenta maior interesse para a gestão acadêmica, pois um FP refere-se a um estudante regular identificado erroneamente como desistente, enquanto um FN é um estudante desistente que foi classificado como regular. Ao analisar as matrizes, observa-se que, nos dados completos, a maior parte das classificações incorretas são FPs (303), enquanto nos dados com os últimos registros, os FNs predominam (59).

Do ponto de vista da gestão, a minimização dos FNs é de extrema importância, pois cada FN representa um aluno evadido que não foi identificado corretamente, o que pode resultar em uma falha no suporte necessário para esses alunos. Em contraste, os FPs, que indicam alunos regulares classificados como evadidos, podem ser vistos como uma oportunidade para a gestão educacional: esses alunos podem estar em risco de evasão no futuro, mas ainda não foram identificados corretamente. Isso sugere que esses estudantes poderiam se beneficiar de monitoramento adicional, suporte acadêmico e intervenções preventivas, com o intuito de evitar sua futura evasão. Uma linha de investigação para trabalhos futuros seria aprofundar o entendimento sobre os fatores que levam ao surgimento de FPs. Estes podem, na verdade, representar alunos que estão em risco de desistir em algum momento do curso, o que poderia indicar a necessidade de uma abordagem proativa de acompanhamento.

Os resultados do teste de Wilcoxon, realizado para comparar os desempenhos dos conjuntos de dados “últimos registros” e “dados completos”, indicaram uma diferença estatisticamente significativa entre os dois cenários, com um valor-p de $0{,}001953125$. Assim, rejeita-se a hipótese nula de que os desempenhos médios dos dois conjuntos são iguais. Observou-se que o conjunto “últimos registros” apresentou uma média de desempenho superior $(0{,}9831)$ em comparação com o conjunto “dados completos” $(0{,}9621)$. Essa diferença pode ser atribuída ao fato de os últimos registros capturarem informações mais recentes e diretamente relacionadas

ao momento de desistência, enquanto os dados completos incluem históricos mais distantes, que podem diluir a relevância das variáveis para a predição do abandono do ensino superior.

6 CONCLUSÃO

Este trabalho teve como objetivo investigar evasão estudantil universitária utilizando dados do sistema acadêmico da UTFPR, campus Dois Vizinhos, e empregar algoritmos de AM para prever estudantes que possam evadir do ensino superior. Após a análise exploratória dos dados e a identificação dos atributos mais relevantes indicados pelo modelo, foi possível destacar as principais características dos alunos que abandonam o curso, identificar o grupo mais propenso a esse comportamento e desenvolver modelos preditivos capazes de classificar a situação dos estudantes com alta precisão.

Os resultados obtidos demonstram que a evasão estudantil é um fenômeno multifacetado, influenciado por uma combinação de fatores acadêmicos, sendo difícil de rotular unicamente por meio de modelos matemáticos e estatísticos. A análise dos dados revelou padrões significativos, como o baixo desempenho acadêmico, que podem ser utilizados para desenvolver modelos preditivos eficazes, que podem ser capazes de identificar estudantes propensos à evasão com alta acurácia.

6.1 Resultados Gerais

Os resultados gerais deste estudo destacam como o AM é promissor ao encontrar padrões nos dados e rotular com alta acurácia e precisão a situação dos estudantes, com o *RF* atingindo valores médios de $0{,}99$ de *F1-Score*. Os modelos se destacam com maior eficiência até o 6º período, onde as chances de evasão são maiores e há muitos exemplos disponíveis para o aprendizado dos mesmos. No entanto, há também a necessidade de um entendimento mais profundo das características que influenciam as taxas de desistência. Muitas das “características relevantes” identificadas não oferecem novos *insights* sobre a situação de desistência, sugerindo a necessidade de um conjunto de dados mais robusto e menos propenso ao *overfitting*, incluindo informações socioeconômicas, histórico de reprovações em disciplinas, frequência às aulas e outros aspectos relacionados.

Observou-se, ao longo dos experimentos, que períodos acadêmicos com menos de 100 exemplos de evasores não justificam grandes esforços para identificação, pois resultam em *folds* de teste extremamente desbalanceados. Essa observação é consistente com a análise apresentada na subseção 4.3.3, onde a 14a ilustra a distribuição das situações dos alunos no *dataset* completo, e a 14b destaca a distribuição levando em consideração apenas o último registro de cada aluno. Portanto, é mais eficaz concentrar os esforços preditivos nos primeiros períodos do curso, onde a evasão é mais expressiva. A partir do sexto semestre, informações adicionais são necessárias para entender melhor as razões da evasão.

Ao utilizar apenas os últimos registros, o modelo alcançou um desempenho ligeiramente superior, com um *F1-Score* de $0{,}997$ na tarefa de "Regular vs Desistente", com o *RF*. Esse resultado reflete uma leve vantagem de se trabalhar com dados mais representativos do sta-

tus acadêmico atual de cada aluno. O registro mais recente de cada aluno fornece uma visão consolidada do progresso acadêmico e da posição atual dentro do curso. O uso de registros históricos de semestres anteriores pode introduzir "preconceitos temporais", como assumir que o simples fato de o aluno estar no 1º período o torna um potencial evasor. Além disso, essa abordagem pode adicionar uma complexidade desnecessária à análise, desviando o foco principal de compreender o status acadêmico mais recente dos alunos.

Parte dos resultados parciais deste trabalho foram apresentados no 35º Simpósio Brasileiro de Informática na Educação (SBIE), realizado em Novembro de 2024, no Rio de Janeiro/RJ. O artigo, intitulado como "*Investigating Student Dropout Risk in Higher Education through Machine Learning*", foi publicado nos anais do evento e está disponível no portal da Sociedade Brasileira de Computação (Parra *et al.*, 2024):

- PARRA, P. et al. ***Investigating student dropout risk in higher education through machine learning***. In: Anais do XXXV Simpósio Brasileiro de Informática na Educação. Porto Alegre, RS, Brasil: SBC, 2024. p. 3020–3028. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/31464

6.2 Limitações e Dificuldades

As limitações deste estudo estão principalmente relacionadas às características presentes nos dados disponíveis. Não foi possível determinar empiricamente todas as razões que levam um aluno a desistir da graduação, especialmente nos períodos mais avançados do curso. Fatores subjacentes, como motivação pessoal e afinidade com o curso, não são capturados pelos dados disponíveis, o que restringe a abrangência das conclusões. Essas variáveis são essenciais para uma compreensão mais profunda dos fatores que influenciam o abandono escolar, mas sua falta limitou a capacidade de desenvolver modelos preditivos mais robustos e abrangentes. Já as dificuldades técnicas enfrentadas durante o processo envolveram a necessidade de um pré-processamento rigoroso dos dados, que demandou um esforço considerável. Esse processo incluiu etapas de limpeza, padronização e checagem manual para garantir a precisão e consistência dos registros utilizados.

6.3 Trabalhos Futuros

Pesquisas futuras devem focar em características que permitam previsões e análises mais precisas, com a realização de uma engenharia de características abrangente, incluindo dados socioeconômicos. Uma abordagem potencial seria buscar a inclusão das disciplinas nas quais o aluno foi reprovado, o que possibilitaria, por exemplo, avaliar a afinidade de um estudante de engenharia com disciplinas matemáticas. Outra variável a ser investigada é a distância entre

a cidade de origem do aluno e o campus universitário, para verificar sua relevância nos modelos preditivos.

Uma linha promissora seria a implementação de estratégias de aprendizado temporal, monitorando as probabilidades de desistência ao longo do tempo. Isso permitiria a identificação de padrões comportamentais que antecedem a evasão, possibilitando intervenções mais eficazes. Idealmente, o sistema acadêmico poderia ser integrado a esses modelos, alertando gestores e profissionais de apoio sobre alunos em risco de evasão, permitindo uma compreensão mais precisa de suas necessidades e facilitando intervenções precoces.

Por fim, técnicas de otimização, podem ser aplicadas para reduzir os erros dos modelos, especialmente os falsos negativos — alunos que desistiram, mas não foram identificados como risco. A utilização dessas técnicas permitirá a seleção eficiente dos melhores modelos, sem a necessidade de buscas exaustivas, agilizando o processo de desenvolvimento e favorecendo a aplicação prática dos resultados.

6.4 Considerações Finais

Este trabalho estabelece uma base sólida para pesquisas futuras na área de predição de evasão escolar e AM. Os modelos desenvolvidos podem ser aprimorados e expandidos por outros pesquisadores, sendo aplicados em diferentes contextos e populações estudantis. Além disso, os resultados obtidos podem auxiliar administradores e educadores a compreender melhor os fatores que contribuem para a evasão, possibilitando a criação de programas de apoio mais eficazes.

O desenvolvimento deste estudo demonstra o potencial do AM como ferramenta estratégica para enfrentar desafios educacionais. Ao identificar alunos em risco de evasão, as instituições podem implementar intervenções personalizadas, como tutoria, aconselhamento e suporte financeiro, promovendo uma experiência acadêmica mais positiva e produtiva. Por fim, este estudo contribui para a compreensão da evasão estudantil e oferece ferramentas práticas para sua previsão e mitigação. Contudo, é importante reconhecer as limitações do estudo, como a dependência de dados históricos e a necessidade de validação contínua dos modelos preditivos em diferentes contextos. Pesquisas futuras podem explorar a integração de novas variáveis e técnicas de AM, visando aprimorar ainda mais a precisão e aplicabilidade dos modelos desenvolvidos.

REFERÊNCIAS

AHMED, S. Comparative study of machine learning techniques in predicting students dropout in colleges of education in north-eastern nigeria. **International Journal of Assessment and Evaluation in Education**, v. 4, n. 8, Apr. 2024. Disponível em: https://mediterraneanpublications.com/mejaee/article/view/376.

AHMED, S. A.; KHAN, S. I. A machine learning approach to predict the engineering students at risk of dropout and factors behind: Bangladesh perspective. *In*: . [*S.l.*]: IEEE, 2019. p. 1–6. ISBN 978-1-5386-5906-9.

ALPAYDIN, E. **Machine Learning**. Cambridge: The MIT Press, 2021. ISBN 9780262365369.

AMARAL, F. **Aprenda Mineração de Dados: Teoria e prática**. Rio de Janeiro: Alta Books, 2016. (Autoria Nacional). ISBN 9788576089889.

BAGGI, C. A. D. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Revista da Avaliação da Educação Superior (Campinas)**, Publicação da Rede de Avaliação Institucional da Educação Superior (RAIES), da Universidade Estadual de Campinas (UNICAMP) e da Universidade de Sorocaba (UNISO)., v. 16, p. 355–374, jul. 2011. ISSN 1414-4077. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772011000200007&lng=pt&tlng=pt.

BARLEM, J. G. T. *et al.* Opção e evasão de um curso de graduação em enfermagem: percepção de estudantes evadidos. **Revista Gaúcha de Enfermagem**, Universidade Federal do Rio Grande do Sul. Escola de Enfermagem, v. 33, p. 132–138, jun. 2012. ISSN 1983-1447. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1983-14472012000200019&lng=pt&tlng=pt.

BARRETO, C. A. d. S. *et al.* Pbil autoens: uma ferramenta de aprendizado de máquina automatizado integrada à plataforma weka / pbil autoens: an automated machine learning tool integrated to the weka ml platform. **Brazilian Journal of Development**, v. 5, n. 12, p. 29226–29242, Dec. 2019. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/5205.

BARROS, B. D. M. **Aprendizado de máquina automático aplicado à predição da evasão no ensino superior**. 2022. 100 p. Dissertação (Mestrado) — Universidade Federal de Goiás, Goiânia, 2022.

BIAZUS, C. A. **Sistema De Fatores Que Influenciam O Aluno a Evadir-Se Dos Cursos De Graduação Na UFSM E Na UFSC: Um Estudo No Curso De Ciências Contábeis**. 2004. 203 p. Tese (Doutorado) — Universidade Federal de Santa Catarina, Florianópolis, 2004.

BISHOP, C. **Pattern Recognition and Machine Learning**. Springer, 2006. (Information Science and Statistics). ISBN 9780387310732. Disponível em: https://books.google.com.br/books?id=kTNoQgAACAAJ.

BRAND C., E. J. *et al.* Towards educational sustainability: An ai system for identifying and preventing student dropout. **IEEE Revista Iberoamericana de Tecnologias del Aprendizaje**, p. 1–1, 2024.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

CARDOSO, F. S. *et al.* O uso da inteligência artificial na educação e seus benefícios: uma revisão exploratória e bibliográfica. **Revista Ciência em Evidência**, v. 4, n. FC, jun. 2023. Disponível em: https://ojs.ifsp.edu.br/index.php/cienciaevidencia/article/view/2332.

CARVALHO, I. L. d. **Reinforcement learning: Um pequeno panorama do Aprendizado Por Reforço**. Slideshare, 2018. Disponível em: https://pt.slideshare.net/slideshow/reinforcement-learning-um-pequeno-panorama-do-aprendizado-por-reforo/105195472.

CASTRO, C. L. de; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, Sociedade Brasileira de Automática, v. 22, p. 441–466, out. 2011. ISSN 0103-1759. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-17592011000500002&lng=pt&tlng=pt.

CASTRO, L. N. de; FERRARI, D. G. **Introdução á Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações**. São Paulo: Saraiva, 2016. ISBN 9788547200992.

CERRI, R.; CARVALHO, A. C. P. de Leon Ferreira de. Aprendizado de máquina: breve introdução e aplicações. **Cadernos de Ciência & Tecnologia**, v. 34, n. 3, p. 297–313, 2017.

COLPO, M.; PRIMO, T.; AGUIAR, M. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. *In*: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 873–884. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/18114.

COUTO, E. **Bias vs. Variância (Parte 2)**. 2013. Disponível em: https://ericcouto.wordpress.com/2013/07/18/bias-vs-variancia-parte-2/. Acesso em: 12 maio 2024.

CUNHA, E. R.; MOROSINI, M. C. Evasão na educação superior: uma temática em discussão. **Revista Cocar**, Belém, v. 7, n. 14, p. 82–89, 2013.

DONNA, G. T. Análise da evasão estudantil com estudo de caso o ifes campus cachoeiro de itapemirim utilizando recursos do machine learning. Trabalho de Conclusão de Curso, Instituto Federal do Espírito Santo, Cachoeiro de Itapemirim, 2023.

EVANGELISTA, R. W. **Estudo da evasão do Bacharelado em Humanidades da UFVJM: causas e consequências**. 2017. 75 p. Tese (Doutorado) — Programa de Pós-Graduação em Educação, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 2017.

FACELI, K. *et al.* **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro: Livros Técnicos e Científicos Editora Ltda - GEN, 2021. ISBN 9788521637493.

FIALHO, M. G. D. **A evasão escolar e a gestão universitária: o caso da Universidade Federal da Paraíbar**. 2014. 107 p. Dissertação (Mestrado) — Universidade Federal da Paraíba, João Pessoa, 2014.

FILHO, F. W. B. H.; VINUTO, T. S.; LEAL, B. C. Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. *In*: **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 1132–1141. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/12869.

FILHO, R. L. L. e S. *et al.* A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, Fundação Carlos Chagas, v. 37, p. 641–659, dez. 2007. ISSN 0100-1574. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-15742007000300007&lng=pt&tlng=pt.

GÉRON, A. **Mãos à obra aprendizado de máquina com Scikit-Learn, Keras & TensorFlow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes**. Traduzido por Rafael Contatori. Rio de Janeiro: Alta Books, 2019. ISBN 9786555208146.

GRUS, J. **Data Science do zero: primeiras regras com o Python**. Traduzido por Welington Nascimento. Rio de Janeiro: Alta Books, 2016. 218 – 221 p. ISBN 9788550803876.

HAN, J.; PEI, J.; TONG, H. **Data Mining: Concepts and Techniques**. [*S.l.*]: Morgan kaufmann, 2022. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780128117613.

JR, P. L.; OSTERMANN, F.; REZENDE, F. Análise dos condicionantes sociais da evasão e retenção em cursos de graduação em física à luz da sociologia de bourdieu. **Revista Brasileira de Pesquisa em Educação em Ciências**, v. 12, n. 1, p. 37–60, ago. 2012. Disponível em: https://periodicos.ufmg.br/index.php/rbpec/article/view/4218.

JúNIOR, R. M. N. *et al.* Educational data mining: PrediÇÃo de evasÃo escolar em colÉgios pÚblicos de salvador. **Apoena Revista Eletrônica**, v. 1, n. 5, p. 29–43, Dez. 2022.

KABATHOVA, J.; DRLIK, M. Towards predicting student's dropout in university courses using different machine learning techniques. **Applied Sciences**, v. 11, n. 7, 2021. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/11/7/3130.

KLOSTERMAN, S. **Projetos de Ciência de Dados com Python: Abordagem de estudo de caso para a criação de projetos de ciência de dados bem-sucedidos usando Python, pandas e scikit-learn**. Traduzido por Aldir C. Côrrea da Silva. São Paulo: Novatec Editora, 2020. ISBN 9786586057102.

KOENIGKAM SANTOS, M. *et al.* Inteligência artificial, aprendizado de máquina, diagnóstico auxiliado por computador e radiômica: avanços da imagem rumo à medicina de precisão. **Radiologia Brasileira**, Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, v. 52, p. 387–396, Nov/Dez 2019. ISSN 1678-7099. Disponível em: https://www.scielo.br/j/rb/a/9yX6w83KDDT33m6G9ddCqBn/?lang=pt#.

LIANG, J. Confusion matrix: Machine learning. **POGIL Activity Clearinghouse**, v. 3, n. 4, Dec. 2022. Disponível em: https://pac.pogil.org/index.php/pac/article/view/304.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, Instituto de Estudos Avançados da Universidade de São Paulo, v. 35, p. 85–94, abr. 2021. ISSN 1806-9592. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142021000100085&tlng=pt.

MANHÃES, L. M. B.; CRUZ, S. M. S. da; ZIMBRÃO, G. Wave: an architecture for predicting dropout in undergraduate courses using edm. *In*: **Proceedings of the 29th Annual ACM Symposium on Applied Computing**. New York, NY, USA: Association for Computing Machinery, 2014. (SAC '14), p. 243–247. ISBN 9781450324694. Disponível em: https://doi.org/10.1145/2554850.2555135.

MANTOVANI, R. G. *et al.* Better trees: an empirical study on hyperparameter tuning of classification decision tree induction algorithms. **Data Mining and Knowledge Discovery**, v. 38, p. 1364–1416, 5 2024. ISSN 1384-5810.

MARSLAND, S. **Machine Learning: An Algorithmic Perspective**. [*S.l.*]: CRC Press, 2011. ISBN 9781420067194.

MARTINS, C. V. M. *et al.* Modelos de previsão de evasão tardia na graduação de uma universidade pública. *In*: . [*S.l.*]: Sociedade Brasileira de Computação - SBC, 2023. p. 41–50.

MATSUBARA, E. T. O algoritmo de aprendizado semi-supervisionado co-training e sua aplicação na rotulação de documentos. mai. 2004.

MATTA, C. M. B. da; LEBRãO, S. M. G.; HELENO, M. G. V. Adaptação, rendimento, evasão e vivências acadêmicas no ensino superior: revisão da literatura. **Psicologia Escolar e Educacional**, Associação Brasileira de Psicologia Escolar e Educacional (ABRAPEE), v. 21, p. 583–591, dez. 2017. ISSN 2175-3539. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-85572017000300583&lng=pt&tlng=pt.

MCCALLUM, A.; NIGAM, K. *et al.* A comparison of event models for naive bayes text classification. *In*: MADISON, WI. **AAAI-98 workshop on learning for text categorization**. [*S.l.*], 1998. v. 752, n. 1, p. 41–48.

MINISTÉRIO DA EDUCAÇÃO. Comissão especial de estudos sobre a evasão nas universidades públicas brasileiras: Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. 1996. Disponível em: http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=24676. Acesso em: 26 abr. 2024.

MITCHELL, T. **Machine Learning**. Nova York: McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. **Sistemas Inteligentes-fundamentos e aplicações**, v. 1, p. 115–139, 2003.

PARRA, P. *et al.* Investigating student dropout risk in higher education through machine learning. *In*: **Anais do XXXV Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2024. p. 3020–3028. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/31464.

PAZ, F.; CAZELLA, S. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. *In*: **Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (WCBIE 2017)**. [*S.l.*: *s.n.*], 2017. p. 624 – 633.

PEDULLA, N. R. d. L. Interpretabilidade de modelos de aprendizagem de máquina no mercado de seguros. Trabalho de Graduação, Universidade Federal de Pernambuco, Recife, out. 2022.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. **arXiv preprint arXiv:2010.16061**, 2020.

RIBEIRO, M. A. O projeto profissional familiar como determinante da evasão universitária: um estudo preliminar. **Revista Brasileira de Orientação Profissional**, Scielo PePsic, São Paulo, v. 6, n. 2, p. 55 – 70, dez. 2005. ISSN 1679-3390. Disponível em: http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1679-33902005000200006&nrm=iso. Acesso em: 19 abr. 2024.

RODRIGUES, J. R. e João Silva e Leonardo Prado e Alex Gomes e R. Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 29, n. 1, p. 1463, 2018. ISSN 2316-6533. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/8107.

SANTOS, R. S. S. dos. **Evasão Escolar Universitária e Estratégias de Intervenções para Retenção do Estudante: Um Estudo de Caso na Universidade Federal de São Carlos**. ago. 2022. Dissertação (Mestrado) — Universidade de São Paulo, ago. 2022.

SARAIVA, D. *et al.* Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática. *In*: **Anais do XXVII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2019. p. 319–333. ISSN 2595-6175. Disponível em: https://sol.sbc.org.br/index.php/wei/article/view/6639.

SILVA, F. C. D.; CABRAL, T. L. D. O.; PACHECO, A. S. V. Evasão em cursos de graduação: Uma análise a partir do censo da educação superior brasileira. *In*: **XVI Coloquio Internacional de Gestión Universitaria - CIGU**. Arequipa - Peru: [*s.n.*], 2016. ISBN 978-85-68618-02-8.

SILVA, J. J. da. **Uma comparação de técnicas de Aprendizado de Máquina para predição de evasão de estudantes no ensino público superior**. mar. 2022. 77 p. Dissertação (Mestrado) — Universidade de São Paulo, São Paulo, mar. 2022.

SILVA, L. da; PERES, S.; BOSCARIOLI, C. **Introdução à Mineração de Dados: Com Aplicações em R**. São Paulo: Elsevier Brasil, 2017. ISBN 9788535284478.

SOLIS, M. *et al.* Perspectives to predict dropout in university students with machine learning. *In*: **2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)**. [*S.l.*: *s.n.*], 2018. p. 1–6.

SONNENSTRAHL, T. S.; BERNARDI, G.; PERTILE, S. Análise de interações do ambiente virtual de aprendizagem para predição de evasão em cursos no ensino a distância. **EaD em Foco**, v. 11, n. 1, jul. 2021. Disponível em: https://eademfoco.cecierj.edu.br/index.php/Revista/article/view/1463.

SOUZA, C. G. *et al.* Algoritmos de aprendizagem de mÁquina e variÁveis de sensoriamento remoto para o mapeamento da cafeicultura. **Boletim de Ciências Geodésicas**, Universidade Federal do Paraná, v. 22, p. 751–773, 12 2016. ISSN 1982-2170. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-21702016000400751&lng=pt&tlng=pt.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction**. [*S.l.*]: MIT press, 2018.

TEODORO, L. de A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. **Revista Brasileira de Informática na Educação**, v. 28, p. 838–863, nov. 2020. ISSN 2317-6121.

VOSSEN, L. V. *et al.* Dropoutless: plataforma colaborativa de predição de evasão. *In*: **Anais do XVIII Simpósio Brasileiro de Sistemas Colaborativos**. Porto Alegre, RS, Brasil: SBC, 2023. p. 193–201. ISSN 2326-2842. Disponível em: https://sol.sbc.org.br/index.php/sbsc/article/view/24234.