Universidade Estadual de Londrina

GUSTAVO MARCELINO DIONISIO

UM PROCESSO BASEADO EM SELEÇÃO DE ATRIBUTOS E APRENDIZAGEM DE MÁQUINA PARA GERAÇÃO DE MODELOS PREDITIVOS: UM ESTUDO SOBRE EVASÃO NO ENSINO SUPERIOR BRASILEIRO

LONDRINA

2024

GUSTAVO MARCELINO DIONISIO

UM PROCESSO BASEADO EM SELEÇÃO DE ATRIBUTOS E APRENDIZAGEM DE MÁQUINA PARA GERAÇÃO DE MODELOS PREDITIVOS: UM ESTUDO SOBRE EVASÃO NO ENSINO SUPERIOR BRASILEIRO

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. André Luís Andrade Menolli

LONDRINA
2024

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

G982u Dionisio, Gustavo Marcelino.
Um processo baseado em seleção de atributos e aprendizagem de máquina para geração de modelos preditivos: um estudo sobre evasão no ensino superior brasileiro / Gustavo Marcelino Dionisio. - Londrina, 2024.
58 f. : il.

Orientador: André Luís Andrade Menolli.
Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, 2024.
Inclui bibliografia.

1. feature selection - Tese. 2. machine learning - Tese. 3. evasão no ensino superior - Tese. I. Andrade Menolli, André Luís. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU 519

GUSTAVO MARCELINO DIONISIO

UM PROCESSO BASEADO EM SELEÇÃO DE ATRIBUTOS E APRENDIZAGEM DE MÁQUINA PARA GERAÇÃO DE MODELOS PREDITIVOS: UM ESTUDO SOBRE EVASÃO NO ENSINO SUPERIOR BRASILEIRO

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

Orientador: Prof. Dr. André Luís Andrade Menolli
Universidade Estadual do Norte do Paraná – UENP

Prof. Dr. Jacques Duílio Brancher
Universidade Estadual de Londrina – UEL

Prof. Dr. Clodis Boscarioli
Universidade Estadual do Oeste do Paraná – UNIOESTE

Londrina, 04 de Abril de 2024.

Agradecimentos

Com profundo apreço, expresso minha gratidão a todos que me apoiaram nesta jornada. Em especial, minha família, cujo suporte incondicional foi essencial. Minha esposa Klaudia, cujo incentivo constante e lembretes sobre a importância dos estudos foram cruciais, nunca permitindo que eu desistisse. Minha filha Laura, luz de alegria que chegou durante meu mestrado e encheu nossos dias mais difíceis de felicidade.

Agradeço também aos meus pais, Carlos e Dalva, e às minhas irmãs, Bruna e Mônica, por acreditarem em mim e fortalecerem minha resiliência e determinação. Um agradecimento especial à liderança da empresa Webdança — Geraldo, Leonel e Fernanda — pelo apoio e pela flexibilidade que me permitiram frequentar as aulas durante o horário de trabalho.

Não posso deixar de mencionar o Dr. André Menolli, meu orientador, cuja orientação perspicaz foi decisiva, especialmente nos momentos em que considerei desistir. Agradeço ao meu amigo e parceiro de mestrado, Rafael, pelo incentivo mútuo, enfrentando desafios e celebrando conquistas juntos.

Finalmente, agradeço a todos que, direta ou indiretamente, contribuíram para minha jornada acadêmica. Cada um de vocês foi vital para o sucesso desta etapa da minha vida.

DIONISIO, G. M. **UM PROCESSO BASEADO EM SELEÇÃO DE ATRIBUTOS E APRENDIZAGEM DE MÁQUINA PARA GERAÇÃO DE MODELOS PREDITIVOS: UM ESTUDO SOBRE EVASÃO NO ENSINO SUPERIOR BRASILEIRO**. 2024. 57f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2024.

RESUMO

Um dos grandes desafios na aprendizagem de máquina em alguns domínios é a alta dimensionalidade de características. Assim, este estudo propõe um processo focado na seleção de atributos e redução de dimensões para aprimorar modelos preditivos, além de ter como uma das saídas os atributos mais relevantes para a predição em questão. O objeto de estudo para a aplicação deste processo é o fenômeno da evasão em Instituições de Ensino Superior brasileiras, com foco especial em cursos presenciais, utilizando dados providos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Com a adoção de técnicas e algoritmos de Aprendizagem de Máquina, o processo visa identificar os atributos mais impactantes na evasão, otimizando a análise preditiva por meio da eliminação de variáveis irrelevantes ao contexto. Este procedimento inclui etapas essenciais, como transformação e balanceamento de dados, seleção de contexto, seleção empírica e algorítmica de atributos, além de possuir etapas iterativas para refinar os modelos preditivos, culminando na geração de modelos de aprendizagem de máquina especializados em contextos específicos. O processo foi aplicado em cinco diferentes contextos do ensino superior brasileiro. Com os resultados destes experimentos, por meio do processo proposto, foi possível gerar modelos preditivos de dimensionalidade reduzida de maior acurácia que os modelos originais. Além disso, comparando com outras técnicas de seleção de atributos os modelos gerados por meio do processo obteve acurácia superior. Com relação ao estudo sobre evasão, foi possível identificar as principais características relacionadas à contextos específicos. Por fim, foi constatado que existe um conjunto de características comum a todos os cenários estudados, que são essenciais na predição da evasão no ensino superior presencial no Brasil.

Palavras-chave: seleção de atributos, aprendizagem de máquina, evasão

DIONISIO, G. M. **A process based on feature selection and machine learning to generate predictive models: an study on higher education dropout**. 2024. 57p. Master's Thesis (Master in Science in Computer Science) – State University of Londrina, Londrina, 2024.

ABSTRACT

One of the major challenges in machine learning in some domains is the high dimensionality of features. Thus, this study proposes a process focused on attribute selection and dimensionality reduction to enhance predictive models, with one of its outputs being the most relevant attributes for the prediction at hand. The object of study for applying this process is the phenomenon of dropout in Brazilian Higher Education Institutions, with a special focus on face-to-face courses, using data provided by the National Institute for Educational Studies and Research. By adopting Machine Learning techniques and algorithms, the process aims to identify the most impactful attributes on dropout, optimizing predictive analysis by eliminating variables irrelevant to the context. This procedure includes essential steps such as data transformation and balancing, context selection, empirical and algorithmic attribute selection, as well as iterative steps to refine predictive models, resulting in the generation of machine learning models specialized in specific contexts. The process was applied in five different contexts of Brazilian higher education. With the results of these experiments, through the proposed process, it was possible to generate predictive models of reduced dimensionality with higher accuracy than the original models. Furthermore, compared to other feature selection techniques, the models generated through the process achieved superior accuracy. Regarding the study on dropout, it was possible to identify the main characteristics related to specific contexts. Finally, it was found that there is a set of common characteristics to all studied scenarios, which are essential in predicting dropout in face-to-face higher education in Brazil.

Keywords: feature selection, machine learning, dropout

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
BI	*Business Intelligence*
CINE	Classificação Internacional Normalizada da Educação
DW	*Data Warehouse*
EAD	Ensino a Distância
ETL	*Extract, Transform and Load*
FS	*Feature Selection*
IDE	*Integrated Development Environment*
IES	Instituições de ensino superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PP	Pontos Percentuais
SA	Seleção de Atributos

CONTEÚDO

1 INTRODUÇÃO

Quando se trabalha com aprendizagem de máquina, uma característica é uma propriedade mensurável individual do modelo de aprendizagem. Por meio de um conjunto de características qualquer algoritmo de aprendizagem de máquina pode realizar a tarefa de classificação [2].

Nos últimos anos, em vários domínios, as características que compõem um modelo cresceram demasiadamente, levando a criar modelos compostos por centenas de variáveis. Assim, como em geral não se sabe quais dessas características são mais relevantes no processo de classificação, usa-se o conjunto todo. Contudo, é inerente a esse processo o risco de que algumas dessas características sejam consideradas irrelevantes ou até mesmo prejudiciais durante o treinamento da máquina.

Para atenuar esse desafio, uma variedade de técnicas e tecnologias foram concebidas visando a redução de variáveis irrelevantes e redundantes. A seleção de atributos, um processo que implica na eliminação criteriosa de variáveis, desempenha um papel fundamental nesse contexto. Além de aprimorar a compreensão dos dados, tal abordagem minimiza a carga computacional, reduz os efeitos adversos da alta dimensionalidade e incrementa o desempenho do preditor. Na literatura são descritos diversos métodos de seleção de atributos, e estes métodos têm como objetivo identificar um subconjunto de variáveis de entrada capaz de descrever de maneira eficaz os dados.

Esse processo almeja minimizar os efeitos adversos do ruído ou de variáveis irrelevantes, garantindo ao mesmo tempo resultados precisos de previsão [3]. Dessa forma, para eliminar características irrelevantes, é necessário empregar um critério de seleção de atributos capaz de avaliar a relevância de cada atributo em relação às classes ou rótulos de saída [2]. Do ponto de vista da aprendizagem de máquina, a inclusão de características irrelevantes pode comprometer a capacidade do sistema em generalizar o modelo com precisão, tanto para os dados de treinamento quanto para novos dados. Vale ressaltar que a eliminação de atributos não implica na criação de novos atributos.

Um domínio em que existem muitas características é o ensino, especialmente quando se busca prever a evasão no ensino superior. Contudo, compreender as causas da evasão não é uma tarefa fácil. Diferentes países fornecem indicadores que têm sido usados por agências governamentais e pesquisadores para medir o número e a taxa de abandono escolar. No entanto, as taxas de abandono escolar por si só podem não ser suficientes para revelar a extensão do problema. É necessário entender e compreender as causas da evasão em diferentes cenários. Por esta razão, esta questão tem sido abordada por vários trabalhos de diferentes países ao longo dos últimos anos, por exemplo,

[4, 5, 6, 7].

Embora no Brasil, ocorra a divulgação anual dos dados do Ensino Superior, que são viabilizadas pelo Censo do Ensino Superior, uma iniciativa promovida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) [8], poucos trabalhos abordam a fundo o fenômeno da evasão no ensino superior de forma abrangente. Nesse contexto, uma variedade de atributos do aluno, que abrangem tanto aspectos sociais quanto acadêmicos, podem estar envolvidas. Ademais, características relacionadas ao curso e à Instituição de Ensino Superior (IES) também podem exercer influência sobre a predição. A abundância desses atributos dificulta a criação de modelos preditivos eficazes para esse domínio.

Estudos destacam a complexidade e a multidimensionalidade da evasão, apontando para fatores acadêmicos, socioeconômicos e individuais que contribuem para ao abandono escolar [9]. Além disso, pesquisas ressaltam a necessidade de considerar variáveis contextuais, como localização geográfica e modelos educacionais, ao analisar a evasão no ensino superior [10].

Diante da complexidade e diversidade de fatores que impactam a evasão, é necessário meios para entender quais características tem mais peso neste processo. Especificamente para o ensino superior brasileiro, a fonte de dados robusta fornecida pelo INEP [11], pode auxiliar nesse entendimento.

Assim, considerando a necessidade da redução de características para composição de modelos preditivos para evasão, além da necessidade do entendimento da evasão em diferentes cenários, e o entendimento das principais características que levam a evasão, este trabalho propõe um processo baseado em seleção de atributos e aprendizagem de máquina para geração de modelos preditivos, que busca não apenas reduzir dimensões de um conjunto de dados, mas também compreender quais fatores e características exercem maior influência em diferentes contextos analisados [12].

O processo é divido em três fases principais. A primeira fase é responsável por gerar um modelo inicial. A segunda fase é responsável por iterar e gerar modelos menores até chegar em um modelo insatisfatório, e a terceira é responsável por realizar iterações até se atingir o menor modelo aceitável, que é o modelo mais reduzido com a acurácia semelhante ao modelo original. Este modelo final, permite uma análise sobre as principais características do modelo. Além disso, para aplicação do processo, foi necessária a execução de etapas de pré-processamento.

Os resultados desta pesquisa promovem reduções de dimensões consideráveis na maioria dos contextos abordados, bem como a manutenção ou melhoria da acurácia do modelo em relação ao modelo inicial. Em uma comparação com outros algoritmos de seleção de atributos, o processo proposto se mostrou equiparável ou superior na maioria

dos contextos.

Além dos resultados computacionais do processo, também há uma discussão sobre o objeto de estudo deste trabalho, que é a evasão no ensino superior. Foi possível identificar as características mais importantes para cada contexto analisado, e observou-se que existem características comuns a todos os contextos e características específicas para cada contexto, verificando a multidimensionalidade deste fenômeno.

1.1 Objetivos

Esta seção discorre sobre o objetivo geral e os objetivos específicos deste trabalho.

1.1.1 Objetivo geral

O objetivo deste trabalho é propor um processo iterativo para a geração de modelos preditivos com dimensões reduzidas por meio da aplicação de técnicas de aprendizagem de máquina e seleção de atributos.

1.1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

- Definir e aplicar uma abordagem de pré-processamento, utilizando os dados oriundos do Censo do Ensino Superior para geração de modelos iniciais de cursos presenciais do ensino superior brasileiro;
- Definir um processo baseado em seleção de atributos e aprendizagem de máquina para geração de modelos preditivos;
- Aplicar o processo proposto em experimentos compostos de conjuntos de dados distintos;
- Analisar os modelos gerados pelo processo proposto por meio dos dados obtidos nos experimentos.

1.2 Organização da dissertação

Este documento está assim organizado:

Capítulo 1: Introdução
Neste capítulo, são estabelecidas as bases para compreender a importância da predição de evasão no ensino superior, apresentando uma visão geral dos objetivos da pesquisa e delineando a estrutura da dissertação.

Capítulo 2: Fundamentação Teórica

Aqui são discutidos os fundamentos teóricos e os conceitos relacionados à predição de evasão no ensino superior, proporcionando uma base para a abordagem metodológica subsequente.

Capítulo 3: Método de Pesquisa

Este capítulo adentra o desenho da pesquisa, o arcabouço metodológico e os métodos de coleta de dados utilizados para investigar a predição de evasão.

Capítulo 4: Processo Proposto

Neste capítulo é detalhado o processo proposto, centrado em técnicas de Aprendizagem de Máquina e Seleção de Atributos para a geração de modelos preditivos com dimensões reduzidas.

Capítulo 5: Resultados e Análise dos Dados

Apresenta os resultados gerados com a aplicação do processo, com dados dispostos em tabelas e figuras e análises preliminares, proporcionando uma visão geral sobre as descobertas e o desempenho dos modelos.

Capítulo 6: Discussão

Aqui são analisados criticamente os resultados dos experimentos, possibilitando uma compreensão mais profunda das implicações e aplicações dos achados.

Capítulo 7: Considerações Finais e Trabalhos Futuros

O capítulo final oferece uma síntese abrangente dos resultados, destacando as contribuições principais, implicações e possíveis direções para pesquisas futuras em relação ao processo e na área da predição de evasão no ensino superior.

2 FUNDAMENTAÇÃO TEÓRICA

A seção de fundamentação teórica fornece o embasamento conceitual relacionado ao tema da pesquisa. Nela, são apresentados os conceitos de aprendizagem de máquina e seleção de atributos, dois conceitos fundamentais para a definição do processo proposto. É importante ressaltar que, embora a proposta seja aplicável a qualquer domínio, os conceitos apresentados estão relacionados ao domínio utilizado para a validação do processo, a evasão no ensino superior.

Esse estilo de escrita visa não apenas apresentar os principais conceitos computacionais aplicados no trabalho, mas também fornecer subsídios para que o leitor compreenda e diferencie a proposta deste trabalho de outras que utilizam aprendizagem de máquina no entendimento da evasão no ensino superior.

Além dos conceitos computacionais, esta seção também aborda os dados educacionais e a evasão no âmbito do ensino superior. Por fim, são apresentados os trabalhos relacionados.

2.1 Aprendizagem de Máquina

O avanço tecnológico tem impulsionado a transformação no campo educacional, em que a aplicação do aprendizagem de máquina se destaca como uma ferramenta com possibilidades de abordar a evasão no ensino superior. Nesta seção, é explorado o conceito de aprendizagem de máquina, seus benefícios, e, posteriormente, são discutidos os principais algoritmos utilizados neste trabalho.

Trabalhos como [13, 12, 14] exploram o papel da aprendizagem de máquina na identificação de padrões e informações a partir de dados, sem exigir uma programação explícita. Esse campo possibilita a construção de modelos capazes de fazer previsões ou tomar decisões com base em dados previamente treinados, sendo essencial para a análise de evasão no ensino superior.

A utilização do aprendizagem de máquina no contexto da evasão no ensino superior pode proporcionar uma abordagem analítica mais precisa e objetiva dos elementos que influenciam esse fenômeno. Algoritmos como os utilizados por [15] na previsão de desistência em Educação a Distância podem identificar padrões sutis nos dados que indicam a probabilidade de um aluno evadir. Dessa forma, as instituições de ensino podem compreender melhor os indicadores de evasão e implementar medidas preventivas e corretivas.

Dentre a gama de algoritmos de aprendizagem de máquina, foram selecionados empiricamente cinco para a construção do processo proposto:*Random Forest*, *k-Nearest*

Neighbors (kNN), *Support Vector Machine* (SVM), *Naive Bayes* e Regressão Logística. Esses cinco algoritmos são amplamente reconhecidos pela sua eficácia na previsão e análise de dados. No entanto, cada um possui suas próprias vantagens e limitações, tornando-os adequados para diferentes tipos de problemas de aprendizagem de máquina. Portanto, no processo proposto, todos os algoritmos são empregados e seus resultados comparados, a fim de identificar o mais eficiente. Nos próximos parágrafos, é apresentado um breve resumo de cada algoritmo.

Random Forest: O algoritmo *Random Forest* consiste em uma técnica de *ensemble learning* que combina várias árvores de decisão para ampliar a precisão e o desempenho das previsões [16]. Ele é aplicável a tarefas de classificação e regressão, demonstrando eficácia na análise de conjuntos de dados complexos e de alta dimensionalidade. A pesquisa conduzida por [17] sobre mineração de dados em sistemas de gerenciamento de cursos (como o *Moodle*, por exemplo) exemplifica a aplicabilidade de algoritmos de aprendizagem de máquina na educação.

k-Nearest Neighbors (kNN): O algoritmo kNN é um modelo supervisionado que classifica novos dados com base nas classes da maioria dos k exemplos mais próximos no espaço de atributos [18]. É especialmente útil para tarefas de classificação, particularmente quando existem padrões de agrupamento. A revisão de [19] sobre mineração de dados educacionais destaca o kNN como uma técnica que pode ser aplicada para identificar grupos de alunos em risco de evasão, possibilitando estratégias de retenção mais eficazes.

Support Vector Machine (SVM): As SVM são algoritmos de aprendizado supervisionado que mapeiam dados para um espaço de alta dimensão, identificando um hiperplano que melhor separa as classes [20]. São eficazes em problemas de classificação, mesmo em conjuntos de dados de alta dimensão. O estudo de [15] exemplifica a aplicação das SVM na previsão de evasão em EAD, demonstrando como essa abordagem pode ser aplicada para identificar alunos em risco.

Naive Bayes: O algoritmo *Naive Bayes* é uma técnica de classificação baseada no teorema de Bayes, pressupondo independência condicional entre os atributos [21]. Essa simplicidade e eficiência tornam o *Naive Bayes* aplicável a conjuntos de dados massivos e relevantes para previsões probabilísticas. Pesquisas, como a de [22], ilustram a utilidade do *Naive Bayes* na predição de desistência em cursos online, destacando a capacidade do algoritmo de analisar padrões de comportamento dos alunos.

Regressão Logística: A Regressão Logística é um modelo estatístico que estima a probabilidade de ocorrência de um evento [23]. Apesar do nome, é comumente utilizado para tarefas de classificação binária, sendo robusto e interpretável. Estudos como o de [24] exploram o uso da Regressão Logística na análise de evasão, demonstrando como ela pode identificar fatores de risco e fornecer *insights* para estratégias de retenção de alunos.

Diferentes algoritmos de Aprendizagem de Máquina vêm sendo aplicados em estudos sobre evasão no ensino superior. Este estudo, portanto, utiliza alguns dos mais relevantes algoritmos, de forma a buscar um melhor resultado para compreender os padrões de evasão.

2.2 Seleção de Atributos

Em contextos de análise de dados, é comum se confrontar com *datasets* extensos, o qual nem todos os atributos contribuem significativamente para o entendimento do fenômeno estudado. Conforme destacado por [25], é comum que conjuntos de dados apresentem uma multiplicidade de variáveis, e entre elas, muitas vezes, existam aquelas que possuem pouca ou nenhuma relevância para a construção de modelos de aprendizagem de máquina.

Por conseguinte, são desenvolvidos algoritmos específicos para identificar e eliminar esses atributos, otimizando o desempenho e a eficiência dos modelos de aprendizagem de máquina. Três categorias de técnicas se destacam para a seleção de atributos: *filter*, *wrapper* e *embedded*. Dentro de cada uma dessas categorias, diversos métodos têm sido propostos na literatura. A Tabela 1 apresenta alguns dos principais métodos existentes em cada uma das categorias.

Os métodos de *filter* na seleção de atributos funcionam independentemente dos algoritmos de aprendizagem de máquina, utilizando critérios estatísticos ou matemáticos para avaliar a importância das características [2]. Subdivisões principais incluem métodos baseados em correlação, que examinam a relação linear entre características e a variável alvo, e métodos baseados em informação mútua, que medem a dependência entre características e a variável alvo, capturando relações não-lineares [3].

Os métodos *wrapper* na seleção de atributos utilizam um modelo preditivo para avaliar o impacto da inclusão ou exclusão de atributos no desempenho do modelo [2]. Essa abordagem permite uma avaliação mais precisa da relevância de cada atributo para a tarefa de modelagem, levando em consideração a interação entre os atributos e o modelo preditivo escolhido. Dessa forma, os métodos *wrapper* buscam otimizar o conjunto de atributos para maximizar a eficiência e eficácia do modelo de aprendizagem de máquina aplicado.

O método *wrapper* pode ser subdividido em busca sequencial e busca heurística [2]. A busca sequencial envolve adicionar ou remover características de forma iterativa, avaliando o impacto de cada mudança no desempenho do modelo. Já a busca heurística, como os Algoritmos Genéticos, adota estratégias de exploração mais amplas do espaço de busca para encontrar combinações ótimas de características, sem necessariamente seguir uma ordem sequencial, o que pode ser mais eficiente em espaços de busca grandes e

Categoria	Método	Descrição
Filter	Correlação	Avalia a correlação linear entre cada atributo e a variável alvo. Rápido e independente do modelo.
	Informação mútua	Mede a dependência mútua entre variáveis. Útil para capturar relações não-lineares.
Wrapper	Busca sequencial	Adiciona ou remove atributos baseando-se na melhoria do desempenho do modelo. Pode ser mais preciso, mas é computacionalmente custoso.
	Algoritmo genético	Utiliza mecanismos de seleção, cruzamento e mutação para encontrar o melhor conjunto de atributos.
Embedded	Random Forest	Realiza a seleção de atributos durante o treinamento do modelo, identificando a importância dos atributos.
	Penalização	Aplica penalidades no modelo de regressão para redução de dimensionalidade, promovendo a seleção de atributos.

Tabela 1 – Algumas das principais categorias de técnicas de seleção de atributos e seus métodos

complexos.

Os métodos *embedded* integram a seleção de atributos como parte do processo de treinamento do modelo de aprendizagem de máquina, ajustando simultaneamente os parâmetros do modelo e selecionando os atributos mais relevantes [26]. Essa abordagem busca oferecer uma eficiência computacional melhorada, pois a seleção de atributos e o treinamento do modelo ocorrem em uma única etapa, evitando a necessidade de avaliações externas repetidas.

As subdivisões dos métodos *embedded* geralmente incluem algoritmos que incorporam penalidades durante o treinamento para realizar a seleção de atributos ou o uso de árvores de decisão, como *Random Forests*, que avaliam a importância dos atributos durante a construção do modelo. Essas técnicas diferem na forma como impõem restrições ou avaliam a importância dos atributos, mas todas visam reduzir a dimensionalidade ao mesmo tempo que mantêm ou até melhoram o desempenho do modelo.

A seleção adequada de atributos é um elemento fundamental no desenvolvimento de modelos de aprendizagem de máquina eficazes, necessitando de uma análise criteri-

osa que leve em conta fatores como simplicidade, estabilidade, e eficácia na redução de variáveis, além da precisão de classificação e requisitos computacionais [2].

Após testar abordagens e analisar estudos relacionados, optou-se pelo algoritmo Boruta [25] para este trabalho. O algoritmo Boruta utiliza o *Random Forest* para determinar a importância das características em um conjunto de dados, aplicando uma comparação entre atributos reais e atributos sombra para selecionar os mais relevantes. Essa técnica é classificada pelos autores como uma abordagem embutida (*embedded*) que integra a seleção de atributos diretamente no processo de modelagem, oferecendo uma maneira objetiva e estatisticamente fundamentada de identificar características significativas.

Portanto, o Boruta se concentra em determinar a relevância dos atributos presentes em um *dataset*, identificando quais são os que realmente desempenham um papel crucial para a análise em questão. O método empregado pelo Boruta envolve a atribuição de pontuações aos atributos, permitindo uma classificação qualitativa desses elementos. Além disso, o pacote Boruta é amplamente utilizado para seleção de atributos [27] e, quando aplicado a um conjunto de dados com um grande número de atributos, tende a proporcionar uma maior precisão [28].

2.3 Dados Educacionais e Evasão no Ensino Superior

Em primeiro momento, esta seção apresenta um panorama sobre os dados do ensino superior do Brasil, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), uma vez que esta fonte de dados serviu como fonte primária para a constituição do *dataset* utilizado no estudo, também sendo utilizada em outros estudos [29, 30]. Além disso, também são apresentados conceitos relacionados à evasão no ensino superior, haja vista que este foi o foco principal do estudo de validação da proposta.

A utilização de dados públicos para o estabelecimento de um panorama do Sistema Educacional é comum em diversos estudos na literatura, como [4, 29, 13]. Portanto, este tipo de análise possibilita o entendimento do estado atual da educação, colaborando na tomada de decisão por parte dos *stakeholders* diretamente relacionados às IES, bem como visar metodologias que facilitem o aprendizado, evitando a saída antecipada de um determinado grupo [31, 13].

Anualmente, INEP disponibiliza os dados relativos à Educação Superior no Brasil. Desde 1995, esses dados têm sido compilados com o propósito de disseminar informações cruciais para o desenvolvimento de políticas públicas que visam aprimorar o cenário educacional do país [11]. A base de dados elaborada pelo INEP é coletada diretamente das Instituições de Ensino Superior (IES) por meio de questionários desenvolvidos pelo pró-

prio instituto. Essa abrangente fonte de informações abarca dados sobre alunos, cursos e as próprias instituições.

Para a devida preparação e análise dos dados, são fornecidas informações adicionais que facilitam o processo de Extração, Transformação e Carga (ETL). Esse suporte compreende desde orientações sobre a abertura dos arquivos em softwares específicos, até os filtros que se originam dos próprios dados coletados [11]. Esses filtros permitem uma análise qualitativa dos dados, possibilitando uma compreensão mais profunda de fatores como a natureza pública ou privada de uma IES, se um curso é ministrado presencialmente ou a distância e a classificação por gênero dos alunos [8].

Um aspecto crucial a ser considerado é a nomenclatura dos cursos e suas respectivas codificações. Até o ano de 2017, os dados e cursos reportados pelo INEP eram codificados segundo os padrões da Organização para a Cooperação e Desenvolvimento Econômico (OCDE). Entretanto, a partir de 2018, essa codificação foi substituída pela Classificação Internacional Normalizada da Educação (CINE), adotada desde então. Assim sendo, para este trabalho em específico, adaptações foram necessárias, e os cursos foram submetidos a um algoritmo simples de adequação para a nova codificação.

Nesse contexto, a análise dos dados provenientes das IES ganha relevância ao se considerar o panorama mais amplo da Educação Superior no Brasil. A utilização dessas informações ricas e diversificadas, em conjunto com análise de dados e aprendizagem de máquina, possibilita uma compreensão mais profunda dos fatores que contribuem para o fenômeno da evasão no ensino superior.

A evasão no ensino superior envolve o abandono de programas educacionais por alunos, independentemente dos motivos que levam a essa decisão [32]. No âmbito do ensino superior, a evasão pode assumir três formas distintas: 1) evasão de curso, quando um aluno abandona um curso em várias circunstâncias; 2) evasão de instituição, quando um estudante deixa a instituição de ensino; e 3) evasão do sistema de ensino superior, indicando a saída, temporária ou permanente, do ensino superior como um todo [4].

As implicações da evasão impactam tanto os alunos evadidos quanto o país como um todo, em termos econômicos e sociais [33]. O fenômeno da evasão é complexo e é influenciado por uma multiplicidade de fatores, que incluem questões pessoais e individuais, aspectos acadêmicos e pedagógicos, bem como a gestão universitária [34].

Entre os principais fatores associados à evasão no ensino superior, destacam-se as condições de estudo na universidade, circunstâncias externas, informações e requisitos de admissão, desempenho acadêmico prévio na educação básica, características individuais dos estudantes e contexto sociodemográfico [35]. Além disso, a análise da evasão deve considerar os contextos individuais dos estudantes e dos cursos, levando em conta aspectos regionais e a área de estudo [36].

A pesquisa sobre evasão no ensino superior tem sido realizada em diferentes regiões e instituições de ensino, buscando compreender as complexidades desse fenômeno. Os estudos específicos e localizados aprofundam a compreensão da evasão no contexto de instituições e cursos específicos, contribuindo para o desenvolvimento de estratégias eficazes de retenção e suporte aos estudantes. Esse comprometimento com a pesquisa sobre evasão reflete a dedicação das IES em aprimorar a qualidade da educação superior oferecida.

Além disso, a pesquisa em evasão no ensino superior é crucial para a criação de políticas e práticas mais eficazes que visam enfrentar esse desafio complexo. Compreender os fatores que influenciam a evasão em diferentes instituições e cursos é essencial para promover uma experiência educacional mais satisfatória e aumentar as taxas de conclusão. A pesquisa nessa área fornece *insights* que podem auxiliar as IES e os formuladores de políticas educacionais a desenvolver estratégias mais eficazes para combater a evasão e promover a conclusão bem-sucedida dos cursos.

2.4 Trabalhos Relacionados

Na literatura, existem diversos estudos que se concentram na análise da evasão de alunos, tanto em âmbito nacional quanto internacional. Esses estudos abordam desde contextos mais específicos, como cursos de uma instituição de ensino superior (IES) em particular, até abordagens mais abrangentes que examinam a situação de um curso em todo o país [37, 4, 38].

Dentre esses estudos, [38] apresenta uma abordagem que utiliza princípios ligados ao conceito de *Data Warehouse* (DW), demonstrando as diferenças entre cursos da área de Computação em geral e cursos de Licenciatura em Computação, utilizando dados do INEP. No estudo [37], por meio de um levantamento realizado na Universidade Federal da Paraíba para o curso de Licenciatura em Computação, na modalidade de Ensino à Distância (EAD), são abordadas as causas de evasão para este curso.

Já no estudo [39], com um conjunto de dados de 2499 registros de alunos da área de computação (Ciência da Computação, Engenharia de *Software* e Sistemas de Informação) da Universidade Federal de Goiás, aborda-se a eficácia dos estudantes em um contexto de cursos com alto índice de evasão. Além de buscar entender as características que contribuem para a evasão, o estudo também analisa as características que contribuem para a retenção e sucesso dos alunos.

Essas análises auxiliam no processo de compreensão do fenômeno da evasão dentro do contexto específico ao qual a Universidade está inserida.

Além das análises de evasão suportadas por DW e *Business Intelligence* (BI), também existem pesquisas que utilizam técnicas de mineração de dados e inteligência

artificial, como a Aprendizagem de Máquina (ML). A Tabela 2 apresenta os principais estudos encontrados na literatura que estudam o fenômeno da evasão por meio de ML, assim como as principais características destes estudos e as pontuações de testes.

Tabela 2 – Resumo das pesquisas anteriores que analisaram as características mais importantes das previsões de evasão no nível do programa de graduação. Adaptado de Vaarma and Li (2024) [1]

Autor	Qt. Dados	Algoritmos	Performance	Importância das características
Berka and Marek (2021)	3339	DT, LR, RF	acurácias em torno de 80%	percentual de vouchers de crédito perdidos no último semestre
Cannistrà et al. (2022)	31071	DT, GLM, RF	AUC de 0.87 a 0.96	créditos acumulados no primeiro ano
Delen (2010)	16066	DT, LR, NN, SVM	precisão de 75% a 87%	horas ganhas divididas por horas registradas, empréstimo estudantil na primavera, média de notas no outono
Djulovic and Li (2013)	7800	DT, NB, NN, indução de regra	precisão de 66% a 74%, recall de 24% a 52%	performance acadêmica
Martins et al. (2023)	4433	EE, RB, RF, SMOTE	f1-scores de 58% a 66%	créditos acumulados, mas varia com o tempo em 3 pontos dentro do primeiro semestre
Matz et al. (2023)	50095	EN, RF	AUC de 0.65 a 0.79	engajamento no aplicativo, média de pontos, etnia
Nagy and Molontay (2023)	6398	CAT	AUC de 0.774	Média do ensino médio, nota em matemática, tempo para entrar em uma IES
Song et al. (2023)	36000	DT, LightGBM, LR, RF, SVM, XGB	precisão de 72% a 83%	número de bolsas, mensalidade, ano de acesso
Yu et al. (2021)	93457	GBT, LR	precisão de 84%, recall de 54%	gênero, universitário de primeira geração, minoria sub-representada e alta necessidade financeira não são importantes
Vaarma and Li (2024)	8813	CAT, NN, LR	precisão de 70% a 90%	créditos acumulados
Teodoro and Kappel (2020)	376746	NB, KNN, DT, RF, NN	acurácia de 80%	atividade extracurricular, idade, carga horária do curso

Em [40], investigou-se a evasão escolar no nível do programa de graduação, utilizando dados de 3.339 alunos. Eles analisaram características demográficas e de transcrição dos alunos, empregando algoritmos de Árvore de Decisão, Regressão Logística e *Random Forest*. Os resultados indicaram acurácias em torno de 80%, destacando a porcentagem de

vouchers de crédito perdidos no último semestre como a característica mais importante para a previsão de evasão.

O estudo conduzido por [41] analisou a evasão no ensino superior em um contexto mais abrangente, considerando dados demográficos, histórico acadêmico e informações sobre os alunos. A pesquisa utilizou algoritmos de Árvores de Decisão e *Random Forest* para prever a evasão. Os resultados mostraram que os escores variaram de 0,87 a 0,96, indicando um bom desempenho na capacidade de previsão da evasão. O estudo destaca a importância de considerar múltiplos aspectos dos alunos ao desenvolver modelos de previsão de evasão no ensino superior.

Utilizando algoritmos como Árvores de Decisão, Regressão Logística, Redes Neurais e SVN, [42] conseguiu-se uma precisão de 75% a 87%. Os resultados destacaram a importância de fatores como horas obtidas divididas por horas registradas, empréstimo estudantil, e média de notas nos semestres de outono e primavera como indicadores cruciais de evasão.

Outro trabalho que empreendeu uma análise abrangente sobre o fenômeno da evasão no ensino superior, adotando abordagens de mineração de dados para investigar fatores determinantes e prever a evasão estudantil é apresentado em [43]. O estudo empregou um conjunto de dados rico em informações demográficas, histórico acadêmico e financeiro dos alunos. Por meio da aplicação de algoritmos como Árvores de Decisão e *Naive Bayes*, o trabalho alcançou precisões que variaram entre 66% e 74%, com taxas de *recall* de 24% a 52%. Esses resultados evidenciam a importância da análise multidimensional na compreensão e previsão da evasão escolar.

No trabalho [44], os autores desenvolveram modelos para prever desempenho acadêmico e evasão usando dados de 4433 alunos, entre 2009 e 2017, de uma universidade politécnica em Portugal. Utilizando cinco algoritmos de aprendizagem de máquina, o estudo destacou a eficácia do *Random Forest*, especialmente ao final do primeiro semestre, alcançando F1-scores entre 58% e 66%.

Em [45] explora-se a previsão da retenção estudantil utilizando dados sociodemográficos e métricas de engajamento via aplicativo em quatro universidades dos EUA, com um total de 50,095 estudantes. Neste estudo, foi demostrado a possibilidade de prever a evasão após o primeiro semestre com desempenho preditivo médio de 78%, apontando variáveis de engajamento comportamental como incrementos significativos à predição além de variáveis institucionais.

Outro estudo foi conduzido por [46] empregando classificadores de aprendizagem de máquina para prever a evasão de estudantes, utilizando dados demográficos e histórico acadêmico de 6.398 alunos de uma universidade húngara. O estudo alcançou uma pontuação AUC de 0.774 na previsão de evasão em um corte de estudantes matriculados

em 2017. Os autores identificaram que a média de notas do ensino médio, as notas de matemática do ensino médio e o número de anos entre a graduação no ensino médio e a matrícula na universidade foram as características mais importantes para prever a evasão de estudantes neste contexto.

Em um trabalho focado em universitários na Coreia do Sul [47], foram utilizados seis classificadores de aprendizagem de máquina para prever a evasão desses estudantes, analisando dados demográficos, histórico escolar e de presença de mais de 36.000 alunos. Descobriu-se que o número de bolsas de estudo, taxas de matrícula e o ano de ingresso são os fatores mais importantes para o conjunto de dados analisado, alcançando uma precisão entre 72% e 83% na previsão de evasão.

No estudo de [48], investigaram a predição de evasão universitária usando dados de 93.457 estudantes de uma universidade estadunidense, aplicando Árvores de Decisão e Regressão Logística. Eles observaram uma acurácia de 83.6% a 83.9%. Identificaram que características como gênero, status de primeira geração de estudantes universitários, minorias sub-representadas e necessidade financeira elevada tiveram menor importância na predição de evasão.

No estudo [1] explorou-se a previsão de evasão no ensino superior utilizando dados de uma universidade finlandesa. Empregando modelos de aprendizagem de máquina, foi evidenciado que atividades em um ambiente virtual de aprendizagem, créditos acumulados e número de cursos reprovados são importantes para prever a evasão, destacando a relevância dos dados do ambiente virtual de aprendizagem neste estudo.

Em [30] estudou-se o fenômeno da evasão escolar no contexto das IES brasileiras, com o objetivo de identificar as características mais determinantes para os alunos desistirem e, assim, tentar prever a possível evasão de outros alunos. Cinco algoritmos de aprendizagem de máquina (*Naive Bayes*, kNN, Árvores de Decisão, *Random Forest* e Redes Neurais) foram aplicados a um conjunto de dados obtidos do INEP com 376 mil registros. Como principal resultado, o estudo indicou que a evasão de estudantes das IES está mais relacionada com a idade, a carga horária total do curso escolhido e com eventual participação em atividades extracurriculares. Os algoritmos *Random Forest* e Redes Neurais apresentaram melhor desempenho em termos de acurácia (80%) e Macro-F1 (79%) como modelos preditivos.

3 MÉTODO DE PESQUISA

A elaboração do seguinte estudo pode ser dividido em quatro etapas (Figura 1), sendo elas: Planejamento Inicial, Fase Exploratória, Desenvolvimento, Análise e Conclusão. Nas seguintes subseções cada uma das fases da pesquisa é detalhada.

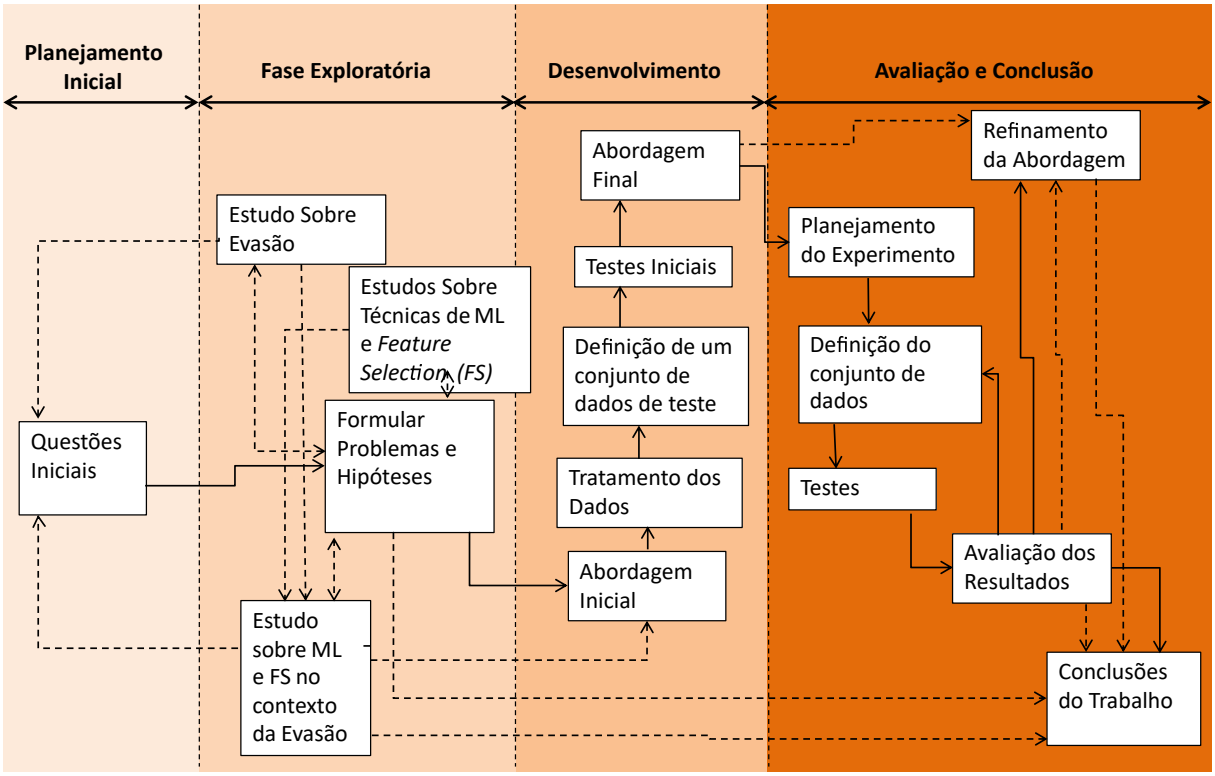

Figura 1 – Metodologia de pesquisa

3.1 Planejamento inicial

A etapa de Planejamento Inicial é constituída principalmente das questões iniciais que motivaram a execução do trabalho, como:

- É possível criar um processo baseado em seleção de atributos o qual se diminua consideravelmente a quantidade de características e gere um modelo satisfatório?
- Esse processo pode obter resultados melhores que os algoritmos existentes de seleção de atributos?
- É possível utilizar esse processo para prever a evasão no contexto do ensino superior?

- O processo consegue ser replicado para ser utilizado em diferentes contexto de análise sobre evasão?
- O contexto em que se analisa a evasão é um fator importante para a acurácia do modelo?
- Utilizando o processo, pode-se entender melhor os atributos que mais impactam a evasão em contextos distintos?

Além das questões iniciais, também foi preciso realizar outras etapas, como análise inicial dos dados públicos do INEP para a compreensão dos tipos e qualidade dos dados.

3.2 Fase exploratória

A Fase Exploratória serviu como base para a investigação em questão. Nesta fase, uma análise preliminar dos dados foi realizada. Além das análises preliminares de dados, nesta etapa também houve revisão da literatura, incorporando principalmente estudos que aplicam técnicas de aprendizagem de máquina e seleção de atributos para abordar problemas de evasão.

O primeiro aspecto essencial da Fase Exploratória foi a revisão de algoritmos de seleção de atributos e revisão de estudos prévios sobre evasão no ensino superior. Isso envolve a investigação de como outros pesquisadores abordaram o fenômeno da evasão, quais variáveis e indicadores foram considerados relevantes e quais métodos foram empregados para analisar os dados educacionais. Essa revisão ajudou a identificar lacunas no conhecimento existente.

Além disso, foi preciso examinar estudos que aplicaram técnicas de aprendizagem de máquina e seleção de atributos para abordar problemas de evasão. Isso permite compreender como o aprendizagem de máquina tem sido utilizado efetivamente para prever e compreender a evasão no ensino superior e indicar as características mais relevantes para o fenômeno. É importante observar quais algoritmos e abordagens têm mostrado sucesso na análise de dados educacionais e como essas técnicas podem ser adaptadas ao contexto da pesquisa em questão.

Outro aspecto importante é a exploração de estudos que se concentram especificamente em técnicas de aprendizagem de máquina. Isso envolve a compreensão de algoritmos, ferramentas e métodos disponíveis para trabalhar com conjuntos de dados. A Fase Exploratória permitiu a identificação de algoritmos que podem ser aplicadas na pesquisa, como *Random Forest*, *kNN*, *SVM*, *Naive Bayes* e Regressão Logística.

Com base na análise da literatura e na compreensão das técnicas de aprendizagem de máquina, pode-se formular problemas de pesquisa e hipóteses. Isso implica na definição

de quais questões serão respondidas pela pesquisa, quais variáveis serão consideradas e quais suposições serão testadas.

Portanto, esta etapa não se limitou apenas à análise preliminar dos dados, mas abrange uma verificação no campo de estudos sobre evasão, estudos que aplicam aprendizagem de máquina e seleção de atributos a esse contexto.

3.3 Desenvolvimento

A fase de Desenvolvimento deste estudo foi dividida em etapas essenciais para a concepção do processo final. A pesquisa começou com a criação de um processo inicial. Este processo foi concebido com base nos *insights* e descobertas da fase exploratória da pesquisa. Foi uma primeira tentativa de estruturar a análise da evasão nas IES, utilizando técnicas de aprendizagem de máquina e seleção de atributos. O objetivo foi criar uma estrutura que pudesse ser testada, validada e, se necessário, adaptada.

A segunda etapa envolveu o tratamento dos dados. Neste momento, foi decidido utilizar o conjunto de dados referente ao ano de 2019, disponibilizado pelo INEP. Para preparar os dados para análise com aprendizagem de máquina, técnicas de transformação de dados foram aplicadas. Isso incluiu a integração dos dados em uma base única, garantindo que estivessem prontos para serem processados de acordo com os requisitos do aprendizagem de máquina.

Com a base de dados preparada, critérios de exclusão foram aplicados para definir um conjunto de dados inicial. Por exemplo, registros que não se relacionavam ao ensino superior foram removidos, uma vez que não estavam alinhados com o objetivo do trabalho. Esse processo de seleção foi fundamental para garantir que o conjunto de dados de teste fosse representativo e relevante para a análise de evasão nas IES.

A etapa de testes iniciais foi realizada para avaliar a eficácia do processo inicial proposto. Durante essa fase, os dados foram alimentados na estrutura concebida, e as técnicas de aprendizagem de máquina foram aplicadas para obter resultados preliminares. Os testes permitiram verificar a viabilidade do processo e identificar áreas que exigiam ajustes.

Com os testes iniciais concluídos e as lições aprendidas, chegou-se à fase de refinamento. O processo inicial foi revisado e modificado com base nas descobertas dos testes. O processo final foi, então, proposto com melhorias e adaptações para melhor atender aos objetivos da pesquisa. Este estágio representa o resultado final da fase de Desenvolvimento, preparando o terreno para a fase de Avaliação e Conclusão.

Em resumo, a fase de Desenvolvimento desta pesquisa envolveu a criação de um processo inicial, o tratamento dos dados, a definição de um conjunto de dados de teste,

testes iniciais para validação e, por fim, a proposição do processo final. Cada uma dessas etapas foi realizada com diligência e propósito, contribuindo para a construção de uma estrutura robusta para a análise da evasão no ensino superior nas IES.

3.4 Avaliação e Conclusão

Na etapa de Avaliação e Conclusão desta pesquisa, foi realizada uma análise dos resultados obtidos após a aplicação do processo proposto em diferentes contextos. Foram utilizadas métricas de desempenho, como acurácia, precisão, *recall* e *F1-score* para avaliar a eficácia dos modelos de previsão de evasão desenvolvidos.

Além disso, foi feita uma comparação do desempenho dos modelos em relação ao número de características utilizadas, observando a importância da seleção de atributos na acurácia das previsões. A partir destes resultados, foram confrontados os achados desse estudo com outros estudos similares encontrados na literatura. Também foram realizados testes comparativos entre os resultados do processo proposto com outras técnicas de seleção de atributos. Por fim, consolidou-se os resultados obtidos para cada contexto.

Essa etapa foi crucial para avaliar a eficácia do processo proposto, além de proporcionar *insights* sobre os principais fatores que influenciam a evasão em diferentes cenários educacionais e validou a utilidade prática da metodologia desenvolvida.

3.5 Materiais e métodos

Esta seção elenca as ferramentas e tecnologias utilizadas na pesquisa, fornecendo uma visão geral das mesmas.

3.5.1 Linguagem R e RStudio

A linguagem R é uma linguagem de programação de código aberto e ambiente de desenvolvimento amplamente utilizado para análise de dados, estatísticas e visualização. Ela é conhecida por sua flexibilidade e pela vasta coleção de pacotes que oferece, tornando-a uma escolha popular entre cientistas de dados e analistas. O *RStudio*, por outro lado, é um ambiente integrado de desenvolvimento (IDE) projetado especificamente para trabalhar com a linguagem R [49].

A combinação da linguagem *R* e do *RStudio* oferece uma plataforma poderosa para análise de dados, modelagem estatística e criação de visualizações informativas. Essas ferramentas são amplamente adotadas em diversas áreas, desde pesquisa acadêmica até aplicações comerciais. Elas capacitam profissionais de dados a explorar, entender e comunicar *insights* a partir de dados de maneira eficaz.

3.5.2 PostgreSQL

O *PostgreSQL* é um sistema de gerenciamento de banco de dados relacional de código aberto amplamente utilizado em todo o mundo. Ele é conhecido por sua robustez, escalabilidade e recursos avançados de *SQL* [50].

Em resumo, o *PostgreSQL* é uma escolha sólida para soluções que buscam um sistema de gerenciamento de banco de dados confiável, escalável e de alto desempenho. Sua conformidade com os padrões SQL, sua extensibilidade e sua ativa comunidade de desenvolvedores fazem dele uma ferramenta versátil para uma variedade de aplicações, desde pequenos projetos até grandes sistemas.

3.5.3 Pentaho Data Integration (Kettle)

A ferramenta *Pentaho Data Integration*, frequentemente chamada de *Kettle*, é uma ferramenta de integração e transformação de dados de código aberto. O *Kettle* é amplamente utilizado para a extração, transformação e carga (ETL) de dados em ambientes de DW e análise de negócios [51].

O *Pentaho Data Integration*, ou *Kettle*, é uma ferramenta amplamente adotada para transformação de dados e ETL. É valioso para quem deseja preparar seus dados para análise, relatórios e tomada de decisões informadas. Sua interface intuitiva e recursos poderosos o tornam uma escolha popular entre profissionais de dados e desenvolvedores de ETL.

3.5.4 Python

Python é a linguagem de programação central empregada para realizar análises e desenvolver modelos de Aprendizagem de Máquina neste estudo. A sua popularidade em ciência de dados e aprendizagem de máquina se deve à sua simplicidade, vasta biblioteca de código aberto e à robusta comunidade de apoio [52].

Utiliza-se amplamente bibliotecas e ferramentas *Python* para diversas tarefas, como manipulação de dados, pré-processamento, treinamento de modelos, avaliação e visualização. As bibliotecas fundamentais incluem *NumPy*, *Pandas*, *Scikit-Learn*, *Matplotlib* e *Seaborn*.

4 PROCESSO PROPOSTO

Antes de aplicar o processo para reduzir as características, é necessário preparar o conjunto de dados e gerar os duas entradas do processo: o modelo inicial e o critério de parada. O modelo inicial delimita o conjunto de características iniciais e as características alvo, enquanto o critério de parada estabelece as métricas para determinar a aceitabilidade de um modelo reduzido.

O fluxo de trabalho para realizar as tarefas de pré-processamento é apresentado na Figura 2. O fluxo de trabalho é dividido em grupos que contêm atividades. As atividades ETL e limpeza dos dados são opcionais e as atividades restantes são obrigatórias.

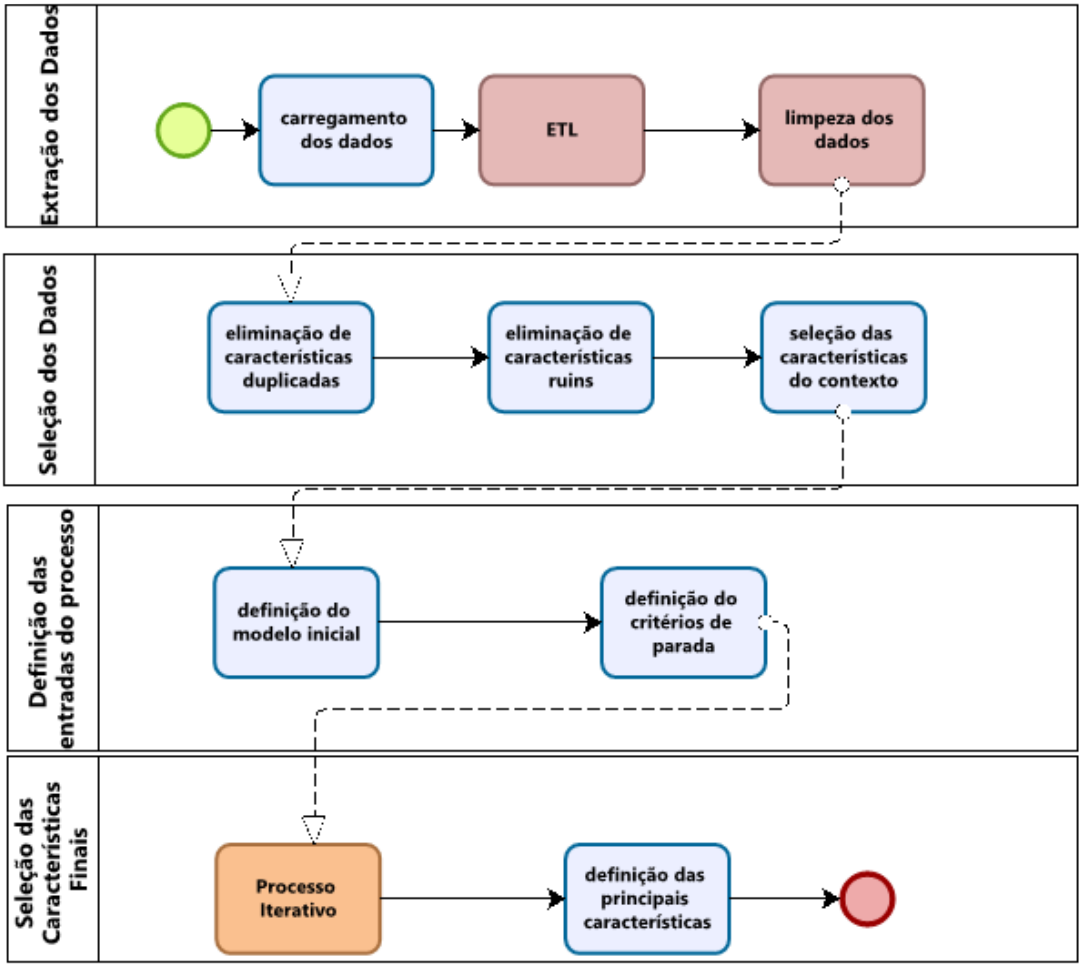

Figura 2 – Fluxo de pré-processamento proposto executado antes do processo

A etapa inicial é o grupo de **extração de dados**. Dentro deste grupo, as atividades englobam o carregamento de dados, a execução de tarefas de extração, transformação e carga conforme necessário, e a realização de tarefas de limpeza de dados.

O segundo grupo é a **seleção de dados**. Inicialmente, envolve a eliminação de atributos duplicados ou altamente semelhantes. Posteriormente, a atenção é direcionada para a remoção de atributos indesejados. Exemplos de tais atributos indesejáveis incluem aqueles com baixa qualidade de dados, inconsistências ou uma quantidade significativa de dados ausentes.

É importante notar que um dos principais fatores que influenciam a análise de evasão é o contexto [36]. Em diferentes contextos, os mesmos atributos podem ter importâncias distintas nos modelos, e o desempenho do modelo tende a melhorar à medida que o contexto é melhor definido. O contexto pode ser definido no escopo de um único curso ou área, abrangendo uma região específica ou mesmo considerando diferentes abordagens de aprendizado, como ensino online ou presencial. De fato, uma das atividades fundamentais dentro deste processo é a definição precisa do contexto e a identificação de seus atributos inicialmente relevantes. Esse passo fundamental lança as bases para as etapas subsequentes do processo.

Uma vez que o contexto é estabelecido, inicia-se o **conjunto de entradas do processo**. Neste grupo, são geradas entradas para o processo (Figura 3). A primeira atividade é definir o modelo inicial. Este modelo inicial serve como base para análises posteriores e é submetido a uma tarefa de classificação.

A última tarefa é definir os critérios de parada. Nesta etapa, é estabelecida a perda máxima de acurácia, precisão, *recall* e F1 aceitáveis em relação ao modelo inicial. Uma vez que vários modelos são gerados com um número reduzido de características, este critério serve para definir quais modelos são aceitáveis.

Uma vez que as entradas são especificadas, o processo para estabelecer modelos com um número reduzido de características é delineado na Figura 3. A etapa inicial (representada em azul mais claro) envolve testar e pontuar o modelo inicial. O processo começa com o cálculo da relevância das características, utilizando o algoritmo de seleção de atributos Boruta [25].

Os resultados desse cálculo orientam as atividades subsequentes destinadas a diminuir o número de características no modelo. Em casos em que certas características permanecem não categorizadas, o algoritmo de correção grosseira tentativa é empregado para facilitar a categorização de características. Em seguida, os dados estatísticos de cada característica, que informam o processo de tomada de decisão do Boruta, são extraídos.

Posteriormente, os dados do modelo inicial são carregados, abrangendo tanto o alvo quanto as características. Em seguida, o método de amostragem é determinado. Neste processo, empregou-se a técnica de validação cruzada k-folds com 10 pastas. Essa técnica é amplamente utilizada em aprendizagem de máquina e estatística para avaliar o desempenho de um modelo preditivo. A validação cruzada k-fold é frequentemente usada

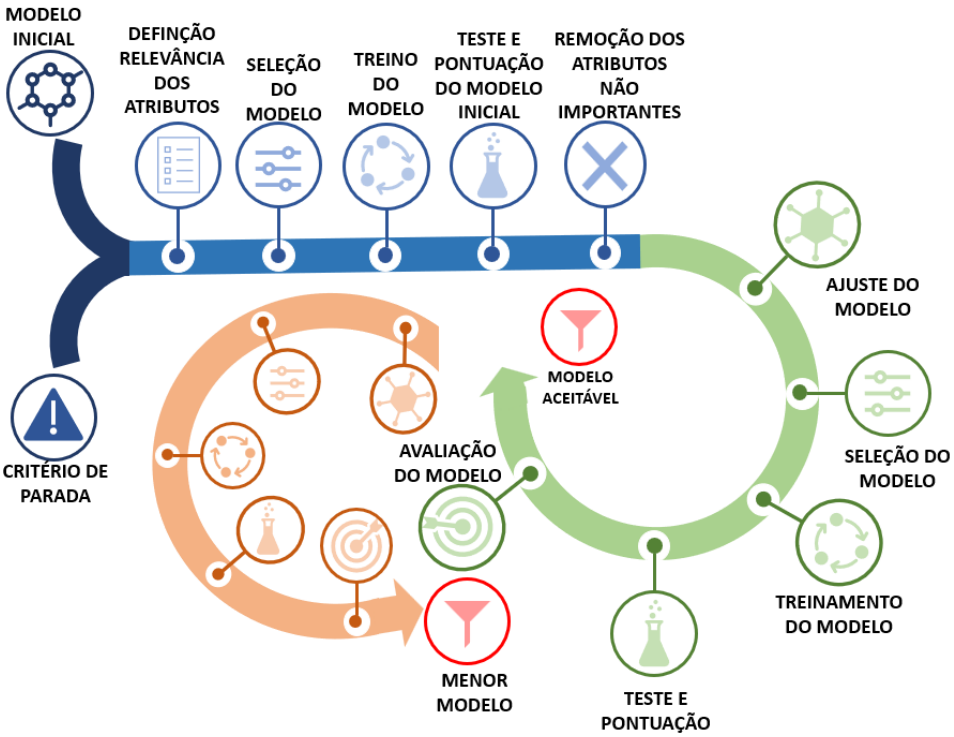

Figura 3 – O processo de definição de novos modelos com número reduzido de funcionalidades.

para ajuste de hiperparâmetros, seleção de modelos e avaliação do desempenho geral de um algoritmo de aprendizagem de máquina. Ajuda a obter uma melhor compreensão de como um modelo generaliza para diferentes subconjuntos de dados e pode fornecer mais confiança nas estimativas de desempenho do modelo.

Após isso, a próxima atividade é a seleção do modelo, o qual os algoritmos são escolhidos. No processo proposto, foram utilizados cinco algoritmos de aprendizagem de máquina: Máquina de Vetores de Suporte (SVM), *Random Forest*, Regressão Logística, K-Vizinhos Mais Próximos (KNN) e Naive Bayes. Esses algoritmos são então aplicados para classificar os dados de treinamento, seguidos pela classificação dos dados de teste, resultando nos respectivos escores de teste. Para pontuar, as seguintes métricas foram utilizadas:

- Área Sob a Curva ROC (AUC): é uma métrica crucial, especialmente para tarefas de classificação binária. Avalia o desempenho de um modelo de classificação medindo sua capacidade de distinguir entre classes positivas e negativas em diferentes configurações de limiar.
- Acurácia: a métrica mais básica e representa a proporção de instâncias corretamente previstas para o total de instâncias. É adequada para conjuntos de dados balanceados, mas pode ser enganosa para conjuntos de dados desbalanceados.

- *Recall*: calcula a proporção de previsões verdadeiramente positivas para o total de positivas reais no conjunto de dados. Avalia o quão bem o modelo identifica instâncias positivas.
- F1 Score: a pontuação é a média harmônica de acurácia e *recall*. Fornece uma medida equilibrada da precisão de um modelo.

A última atividade nesta etapa envolve a remoção de características não importantes, levando em consideração os dados estatísticos do Boruta. Posteriormente, um novo modelo é gerado e utilizado como entrada para o ciclo iterativo (parte verde).

O ciclo iterativo inicia executando o ajuste do modelo, onde 10% das características menos importantes são removidas. Para a primeira iteração, este passo não é realizado. Após isso, os algoritmos são escolhidos, e o modelo é treinado e testado, gerando pontuações.

A etapa final do ciclo iterativo envolve comparar as pontuações do modelo atual com as do modelo inicial. Para decidir se deve prosseguir com iterações adicionais, a avaliação inclui as seguintes métricas: AUC, Acurácia, Precisão, *Recall* e F1 *Score*. Se a diferença entre as métricas for menor que o limite estabelecido nos critérios de parada, o modelo é considerado aceitável, e uma nova iteração é iniciada.

Assim, dentro do ciclo iterativo, cada iteração produz um modelo aceitável com um número reduzido de características até que um modelo inaceitável seja gerado, levando à interrupção de iterações adicionais.

O ciclo final (parte laranja) recebe como entrada o modelo inaceitável. Em cada iteração, a característica mais importante previamente removida é reintroduzida no modelo inaceitável. O modelo é então treinado, testado e comparado com o modelo inicial. Se o modelo atender aos critérios de parada, ele se torna o modelo final; caso contrário, o ciclo se repete. Assim, o ciclo final é repetido até obter um modelo satisfatório, que é o menor modelo.

4.1 Configuração e aplicação do processo

Nesta seção é apresentado como foi feita a configuração do processo proposto e como a mesmo foi aplicado em um dos cinco contexto usados neste trabalho, o contexto de todos os cursos presenciais no Brasil. Para todos os outros contextos utilizados como experimento deste trabalho, seguiu-se a mesma dinâmica descrita nesta seção.

No estudo, o conjunto de dados do censo da educação do ensino superior do ano de 2019, provido pelo INEP [11] foi utilizado. O censo do ano 2019 foi escolhido pois possui dados individualizados dos alunos, enquanto os censos subsequentes tiveram uma

mudança na estrutura dos dados, removendo a individualidade dos alunos e agrupando os dados por curso. Dessa forma, os dados de 2019 foram os dados mais recentes com informações individualizadas dos alunos.

No conjunto de dados original, o arquivo do aluno contém 105 características, o curso 112 e a instituição 48. Este conjunto de dados abrange vários tipos de programas de ensino superior no Brasil, incluindo cursos de graduação e tecnológicos.

No entanto, para este estudo, foram considerados exclusivamente cursos de graduação, concentrando a análise neste subconjunto específico para alinhar-se com o foco de pesquisa.

Na fase de extração de dados, começou-se importando os arquivos para um banco de dados e realizando tarefas de ETL (Extração, Transformação, Carga) e limpeza de dados. No decorrer dessas atividades, foi realizado o seguinte:

1. Integração de Dados: combinação de dados de várias fontes em uma única tabela, consolidando um conjunto de dados para análise simplificada.

2. Ampliação de Dados Externos: Para aprimorar o conjunto de dados, foram incorporadas informações de bancos de dados externos, como o Instituto Brasileiro de Geografia e Estatística (IBGE) para dados georreferenciais, e a Classificação Internacional Padronizada de Educação Adaptada para Cursos de Graduação e Sequências de Formação Específicas (Cine Brasil) para *insights* abrangentes sobre áreas de curso.

Na fase de seleção de dados, características duplicadas ou ruins (muitos dados nulos, sem relação com o objeto de estudo) foram eliminadas, resultando, em última análise, em um conjunto de dados refinado composto por 61 características essenciais. Em seguida, foi definido o contexto para o estudo inicial, que envolveu a análise de cursos de graduação presenciais em todo o país. Dentro desse contexto, o conjunto de dados foi simplificado para consistir em 50 campos.

Além disso, foi elaborado um conjunto de dados balanceado, totalizando 400 mil registros, distribuídos igualmente entre estudantes que abandonaram o curso e aqueles que não o fizeram.

Para a seleção inicial de características, o primeiro passo foi definir o modelo inicial, onde a evasão foi definida como o alvo e os outros 49 campos como características.

Após isso, foi definido o critério de parada. Uma regra restritiva para avaliar a aceitabilidade de um modelo foi estabelecida. Esse critério exigia que nenhuma métrica apresentasse uma deterioração de mais de 2% em comparação com as métricas obtidas a partir do modelo inicial. Portanto, estabeleceu-se um critério onde a diferença máxima

permitida entre o modelo atual e o modelo inicial para todas as métricas é de 2%. Esse limite orienta a avaliação do desempenho do modelo.

Tabela 3 – Resultados após a definição da relevância dos atributos.

Atributo	**média**	**mediana**	**min**	**máx**	**normHits**	**decisão**
tp_modalidade_ensino	0.00	0.00	0.00	0.00	0	Rej.
tp_nivel_academico	0.00	0.00	0.00	0.00	0	Rej.
co_pais_origem	2.16	2.13	-0.63	5.21	0.34	Rej.
in_mobilidade_academica	5.45	5.48	3.68	6.76	1	Confir.
tp_atributo_ingresso	7.97	7.93	5.93	10.08	1	Confir.
tipo_reserva	11.26	11.29	10.03	12.65	1	Confir.
in_reserva_vagas	11.87	11.84	10.73	13.22	1	Confir.
tp_nacionalidade	17.41	17.42	14.86	20.17	1	Confir.
in_ajuda_deficiente	19.50	19.46	16.48	22.93	1	Confir.
in_gratuito	24.55	24.55	23.25	26.02	1	Confir.
in_servico_internet	23.13	23.16	19.64	27.04	1	Confir.
in_bolsa	26.62	26.57	24.29	28.61	1	Confir.
tp_grau_academico	23.48	23.21	17.39	28.83	1	Confir.
in_possui_laboratorio	30.75	30.48	24.69	35.83	1	Confir.
tp_turno	34.11	33.94	32.03	36.42	1	Confir.
in_deficiencia	29.78	29.91	25.30	36.75	1	Confir.
tp_escola_conclusao_ens_medio	32.16	31.94	28.71	37.00	1	Confir.
in_apoio_social	36.56	36.65	33.81	39.59	1	Confir.
tp_organizacao_academica	37.45	37.40	34.44	41.05	1	Confir.
in_assina_outra_base	36.61	36.34	32.21	41.31	1	Confir.
in_disciplina_libras	35.73	35.64	31.69	41.81	1	Confir.
tp_cor_raca	38.94	39.08	34.59	43.81	1	Confir.
in_catalogo_online	40.08	40.15	34.90	44.24	1	Confir.
in_tradutor_libras	42.71	42.60	39.77	45.86	1	Confir.
co_grande_area	40.80	40.81	35.67	46.09	1	Confir.
in_oferece_disc_semi_pres	43.21	43.01	40.86	46.52	1	Confir.
tp_categoria_administrativa	43.61	43.69	41.21	46.62	1	Confir.
in_capital	42.71	42.49	35.44	48.39	1	Confir.
co_cine_area_geral	42.72	42.53	35.55	49.58	1	Confir.
in_acesso_portal_capes	47.24	47.07	43.94	50.33	1	Confir.
tipo_ingresso	44.69	44.42	38.70	51.46	1	Confir.
area	46.99	46.74	40.63	52.46	1	Confir.
tempo_curso_range	47.56	47.27	41.53	53.38	1	Confir.
no_curso	46.16	45.63	40.06	53.51	1	Confir.
co_regiao	44.68	44.80	36.43	54.56	1	Confir.
in_ingresso_total	50.97	50.93	47.08	55.23	1	Confir.
cargar_horaria_curso	50.58	50.65	46.13	55.30	1	Confir.
in_repositorio_institucional	54.04	53.86	48.05	59.00	1	Confir.
co_uf	51.28	50.88	42.05	61.26	1	Confir.
idade	58.20	58.48	54.30	62.37	1	Confir.
in_busca_integrada	58.76	58.91	53.48	62.72	1	Confir.
co_cine_rotulo	57.60	57.58	50.48	63.91	1	Confir.
in_atividade_extracurricular	60.98	60.86	57.24	64.72	1	Confir.
in_ingresso_processo	74.81	74.72	70.04	80.38	1	Confir.
co_municipio	73.00	73.12	64.85	82.32	1	Confir.
in_concluinte	79.59	79.70	75.36	84.22	1	Confir.
co_ies	78.87	78.96	70.23	85.84	1	Confir.
tempo_entrada_curso	89.46	89.73	81.42	98.17	1	Confir.
porc_concluida_range	101.67	101.49	94.44	106.44	1	Confir.

Após definir as duas entradas (a parte reta, em azul escuro do processo na Figura 3), procedeu-se com as etapas subsequentes representadas pela parte azul clara.

A primeira etapa dentro desta fase é a definição de relevância de atributos, que gera uma lista de atributos relevantes. Para gerá-la, foi empregado o algoritmo Boruta no

modelo. O algoritmo Boruta executou 99 iterações, identificando 46 atributos confirmados como importantes e 3 atributos confirmados como irrelevantes. Um resumo das estatísticas de atributos deste subprocesso pode ser encontrado na Tabela 3.

Em seguida, foi executada a seleção do modelo, onde foram empregados cinco algoritmos de ML. Posteriormente, foram realizadas as etapas de treino e teste do modelo inicial, conforme resultados apresentados na Tabela 4.

Tabela 4 – Pontuação do modelo inicial.

	Média total das classes					**Classe Evasão**		
Modelo	AUC	CA	F1	Prec	Recall	F1	Prec	Recall
Random Forest	0.91	0.84	0.84	0.84	0.84	0.87	0.84	0.91
Regressão Logística	0.52	0.61	0.46	0.37	0.61	0.76	0.61	1.00
SVM	0.50	0.61	0.46	0.68	0.61	0.76	0.61	1.00
Naive Bayes	0.73	0.68	0.67	0.67	0.68	0.75	0.72	0.78
kNN	0.81	0.75	0.75	0.75	0.75	0.81	0.76	0.86

Posteriormente, a iteração definida no ciclo iterativo (parte verde) do processo da Figura 3 foi realizada. Este ciclo cria novos modelos com um número reduzido de características, e suas pontuações são então comparadas com as do modelo inicial. Se a diferença entre as métricas cair abaixo do limite definido nos critérios de parada, o ciclo persiste.

Neste ciclo, a primeira iteração remove todas as características que foram sinalizadas como rejeitadas pelo algoritmo Boruta. Em seguida, em cada iteração subsequente, ocorre a remoção de 10% das características menos importantes. Dado que há um total de 49 características, optou-se por remover 5 características durante cada iteração.

Uma vez que o ciclo iterativo (ciclo verde) produz um modelo insatisfatório, este modelo é usado como entrada para o ciclo final (parte laranja) (Figura 3). Neste ciclo, as características mais importantes devem ser reintegradas ao modelo uma por uma, até alcançar um modelo aceitável, ou seja, o menor modelo gerado.

A Tabela 5 exibe um resumo abrangente dos resultados obtidos em cada iteração. Para aumentar a clareza, apenas os melhores resultados do modelo de cada iteração são apresentados, com o *Random Forest* consistentemente apresentando o melhor desempenho em todas as instâncias. A iteração inicial é gerada pela parte azul clara do processo representado na Figura 3. Iterações subsequentes, da primeira (Boruta) à oitava, são fornecidas pelo ciclo iterativo (verde). A oitava iteração resultou em um modelo insatisfatório, destacado na Tabela 5 juntamente com a métrica correspondente. Esta iteração serve de entrada para o ciclo final (laranja), que produz o menor modelo.

É importante destacar que, ao analisar os resultados, isso ocorreu em dois aspectos. Em primeiro lugar, avaliou-se o modelo em sua totalidade, considerando as métricas em todas as classes. Essa comparação teve como objetivo avaliar a qualidade geral do modelo,

ou seja, quão bem ele se saiu na previsão de casos de evasão e não evasão. No entanto, como o objeto deste estudo é a evasão, também foi feita a comparação *recall*, F1 e acurácia para a classe de evasão.

Tabela 5 – Pontuações de teste para todas as iterações do contexto de todos os cursos

	Média total das classes					**Classe Evasão**		
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall
Inicial	0.908	0.839	0.837	0.839	0.839	0.873	0.840	0.909
1(Boruta)	0.908	0.839	0.837	0.839	0.839	0.873	0.839	0.910
2	0.907	0.839	0.837	0.839	0.839	0.873	0.841	0.906
3	0.903	0.835	0.833	0.835	0.835	0.869	0.841	0.900
4	0.898	0.832	0.830	0.831	0.832	0.867	0.839	0.896
5	0.899	0.834	0.832	0.833	0.834	0.868	0.841	0.898
7	0.897	0.833	0.831	0.832	0.833	0.867	0.841	0.894
9	0.891	0.826	0.824	0.825	0.826	0.861	0.837	0.887
Menor	0.897	0.833	0.831	0.832	0.833	0.867	0.842	0.893

5 RESULTADOS E ANÁLISE DOS DADOS

Nesta seção, os resultados dos experimentos conduzidos utilizando o processo proposto são apresentados. Os experimentos foram planejados com dois objetivos principais: (1) Avaliar a eficácia do processo em diversos contextos; (2) Obter *insights* sobre as principais características que influenciam na predição de evasão em diferentes cenários. Considerando isso, o processo foi empregado em cinco contextos distintos.

O contexto inicial refere-se a cursos de graduação presenciais no Brasil. No segundo contexto, todos os cursos presenciais de computação no Brasil foram considerados, abrangendo mais de 15 cursos diferentes. Os três contextos restantes focam cursos de enfermagem. No terceiro contexto, dados dos cursos presenciais de enfermagem em todo o Brasil são explorados. O quarto contexto é uma extensão especializada, concentrando-se em IES privadas no estado de São Paulo. Por fim, o quinto contexto explora os cursos de enfermagem em IES públicas em São Paulo. A Tabela 6 resume as principais informações sobre cada contexto e o conjunto de dados utilizado.

Tabela 6 – Informações sobre os contextos e conjuntos de dados utilizados nos experimentos.

Informações Contexto					**Informações Conjunto de Dados**		
Contexto	**Região**	**Nº Cursos**	**Nº IES**	**Tipo IES**	**Qt. Dados**	**Evadidos**	**Não evadidos**
A	Brasil	251	2381	Púb. e Priv.	400.000	200.000	200.000
B	Brasil	15	661	Púb. e Priv.	72.000	36.000	36.000
C	Brasil	1	904	Púb. e Priv.	72.000	36.000	36.000
C1	São Paulo	1	144	Priv.	40.000	20.000	20.000
C2	São Paulo	1	13	Púb.	700	350	350

Com relação ao Contexto A, a Tabela 7 mostra as métricas obtidas pelo modelo inicial (linha 1) e as diferenças percentuais nas métricas para cada iteração em comparação com as métricas do modelo inicial. A tabela também apresenta o número de características utilizadas em cada iteração.

Considerando os resultados apresentados na Tabela 7, percebe-se que na oitava iteração, o *Recall* da classe evasão apresenta uma variação superior a 2% do modelo inicial. A última iteração fornece a diferença de desempenho entre o o menor modelo e o modelo inicial para cada métrica. Neste contexto, a acurácia do modelo com 13 características foi

apenas 0,75% menor do que a do modelo inicial.

Tabela 7 – A diferença (%) da performance entre o melhor modelo em cada iteração e o modelo inicial para todos os cursos presenciais do Brasil. A coluna QC indica a quantidade de características

	Média total das classes					**Classe Evasão**			
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall	**QC**
Inicial	0.908	0.839	0.837	0.839	0.839	0.873	0.840	0.909	49
1 (boruta)	0.00%	-0.01%	0.00%	-0.02%	-0.01%	-0.02%	0.08%	-0.14%	46
2	0.13%	0.06%	0.05%	0.09%	0.06%	0.08%	-0.12%	0.30%	41
3	0.05%	0.01%	-0.01%	0.03%	0.01%	0.04%	-0.17%	0.27%	36
4	0.53%	0.46%	0.42%	0.53%	0.46%	0.43%	-0.07%	0.97%	31
5	1.03%	0.83%	0.79%	0.92%	0.83%	0.73%	0.11%	1.40%	26
6	0.91%	0.58%	0.53%	0.65%	0.58%	0.54%	-0.07%	1.18%	21
7	1.14%	0.74%	0.67%	0.83%	0.74%	0.70%	-0.15%	1.59%	16
8	1.89%	1.58%	1.51%	1.69%	1.58%	1.36%	0.41%	2.36%	11
Menor	1.19%	0.76%	0.68%	0.86%	0.76%	0.73%	-0.24%	1.77%	13

Outra análise realizada consistiu na comparação entre os resultados obtidos pelo processo proposto e os de outros algoritmos de redução de atributos das categorias apresentadas na Seção 2.2. Para isso, o melhor modelo obtido no Contexto A foi comparado com os algoritmos *Information Gain*, *Chi2*, *GINI*, *ReliefF* e Correlação [53]. Os seguintes passos foram seguidos:

1. Criar um modelo reduzido utilizando o mesmo número de características do melhor modelo gerado pelo processo proposto.
2. Aplicar o algoritmo *Random Forest* utilizando as mesmas configurações de treinamento e teste usadas para gerar os resultados do processo proposto.
3. Comparar os resultados obtidos com os do processo proposto.

A Figura 4 apresenta a comparação das métricas de acurácia, F1 de todas as classes e F1 das classes de evasão. Para o Contexto A, os resultados produzidos pelo modelo do processo foi melhor ou igual cinco dos seis modelos comparado. Contudo, o modelo produzido pelo algoritmo *ReliefF* foi ligeiramente superior.

Para o segundo contexto (B), cursos presenciais de ciência da computação no Brasil, a Tabela 8 apresenta os resultados do experimento. Neste contexto, obteve-se um modelo que exibe uma diminuição de 1,34% na acurácia em comparação com o modelo inicial, enquanto utiliza apenas 28,2% das características iniciais. Essa redução no número de características em quase 70% nos permitiu obter um modelo aceitável, considerando os critérios estabelecidos.

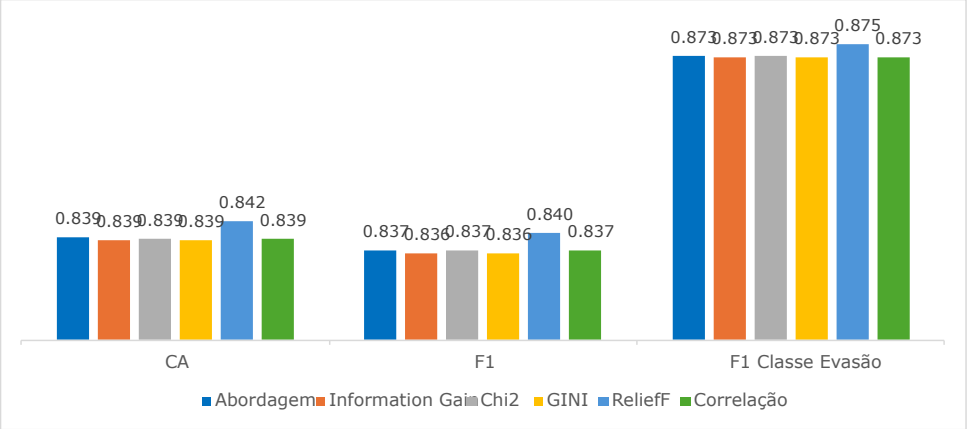

Figura 4 – Comparação entre métricas produzidas pelo melhor modelo para o Contexto A com outros algoritmos de redução de características

Tabela 8 – A diferença (%) da performance entre o melhor modelo em cada iteração e o modelo inicial para o contexto de cursos de computação. A coluna QC indica a quantidade de características.

	Média total das classes					Classe Evasão			
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall	**QC**
Initial	0.844	0.764	0.764	0.764	0.764	0.767	0.758	0.776	47
1 (boruta)	-0.20%	-0.21%	-0.22%	-0.21%	-0.21%	-0.17%	-0.31%	-0.01%	40
2	-0.33%	-0.38%	-0.38%	-0.38%	-0.38%	-0.38%	-0.35%	-0.41%	35
3	-0.18%	-0.24%	-0.24%	-0.24%	-0.24%	-0.23%	-0.25%	-0.21%	30
4	0.07%	0.02%	0.02%	0.02%	0.02%	0.00%	0.07%	-0.08%	25
6	0.53%	0.57%	0.57%	0.55%	0.57%	0.43%	0.85%	0.00%	20
6	1.11%	0.90%	0.90%	0.87%	0.90%	0.70%	1.27%	0.10%	15
7	2.08%	1.80%	1.81%	1.75%	1.80%	1.44%	2.42%	0.42%	10
Menor	1.57%	1.34%	1.35%	1.30%	1.34%	1.07%	1.82%	0.28%	13

Com relação à comparação do modelo produzido pelo processo com outros algoritmos de redução de características, para o Contexto B, todos os algoritmos geraram modelos que produziram resultados quase idênticos, sem diferença significativa entre nenhum modelo, conforme mostrado na Figura 5.

O terceiro contexto (C) diz respeito aos cursos de enfermagem presenciais em todo o Brasil. A Tabela 9 resume os resultados do experimento para este contexto. O menor modelo exibiu ligeiros decrementos de até 0,5% em seis métricas, melhorias em uma métrica e uma única métrica que deteriorou em mais de 1% em relação ao modelo inicial. No entanto, é crucial destacar que este modelo opera com apenas 17% das características iniciais, representando uma redução de quase 85%. Apesar dessa redução significativa de características, conseguiu-se um modelo aceitável, seguindo os critérios estabelecidos.

Com relação à comparação do modelo produzido pelo processo com outros algorit-

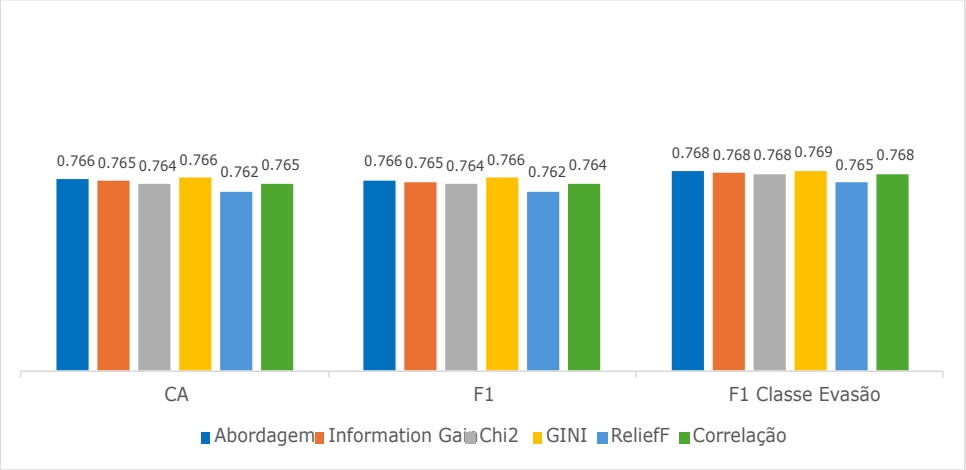

Figura 5 – Comparação entre métricas produzidas pelo melhor modelo para o Contexto B com outros algoritmos de redução de características.

Tabela 9 – A diferença (%) da performance entre o melhor modelo em cada iteração e o modelo inicial para o contexto de cursos de enfermagem no Brasil. A coluna QC indica a quantidade de características.

	Média total das classes					**Classe Evasão**			
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall	**QC**
Inicial	0.883	0.808	0.808	0.808	0.808	0.810	0.805	0.815	47
1 (boruta)	0.02%	0.25%	0.25%	0.25%	0.25%	0.25%	0.24%	0.27%	39
2	-0.02%	0.14%	0.14%	0.14%	0.14%	0.11%	0.25%	-0.03%	34
3	-0.01%	0.12%	0.12%	0.12%	0.12%	0.10%	0.20%	0.00%	29
4	0.11%	0.29%	0.29%	0.28%	0.29%	0.21%	0.52%	-0.11%	24
5	0.32%	0.46%	0.46%	0.45%	0.46%	0.33%	0.87%	-0.23%	19
6	-0.36%	-0.25%	-0.24%	-0.30%	-0.25%	-0.56%	0.78%	-1.96%	14
7	0.03%	0.03%	0.04%	-0.05%	0.03%	-0.38%	1.33%	-2.18%	9
8	1.86%	1.60%	1.61%	1.52%	1.60%	1.17%	2.76%	-0.50%	4
Menor	0.58%	0.58%	0.59%	0.52%	0.58%	0.22%	1.66%	-1.29%	8

mos de redução de características, para o Contexto C, o modelo produzido pelo processo demonstrou resultados superiores a todos os outros algoritmos em todas as métricas comparadas, como ilustrado na Figura 6.

O quarto contexto (C1) é um contexto especializado dentro do campo dos cursos de enfermagem, focando exclusivamente em instituições de ensino superior privadas, com o escopo geográfico limitado ao estado de São Paulo. A Tabela 10 resume os resultados do experimento. O menor modelo exibiu pequenos decrementos de menos de 1,0% em seis métricas, melhorias em uma métrica e uma única métrica que deteriorou em mais de 1%. No entanto, é crucial destacar que este modelo operava com apenas 18,6% das características iniciais, representando uma redução de mais de 80%. Apesar dessa redução

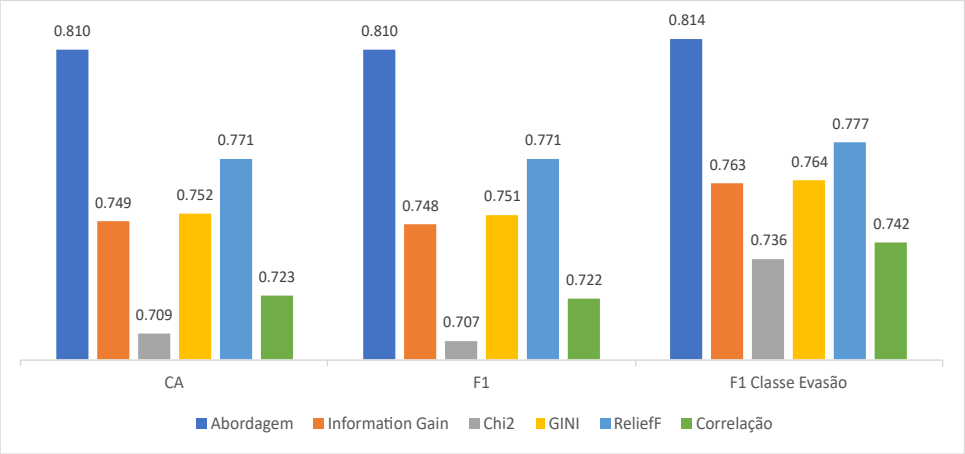

Figura 6 – Comparação entre métricas produzidas pelo melhor modelo para o Contexto C com outros algoritmos de redução de características

significativa de características, foi possível obter um modelo aceitável, seguindo os critérios estabelecidos.

Tabela 10 – A diferença (%) da performance entre o melhor modelo em cada iteração e o modelo inicial para o contexto de cursos privados de enfermagem no estado de São Paulo. A coluna QC indica a quantidade de características.

	Média total das classes					**Classe Evasão**			
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall	**QC**
Initial	0.878	0.802	0.802	0.802	0.802	0.804	0.794	0.814	43
1 (boruta)	-0.21%	-0.11%	-0.11%	-0.11%	-0.11%	-0.06%	-0.25%	0.13%	34
2	-0.22%	-0.32%	-0.32%	-0.33%	-0.32%	-0.34%	-0.25%	-0.44%	30
3	-0.31%	-0.42%	-0.42%	-0.42%	-0.42%	-0.43%	-0.35%	-0.52%	26
4	-0.33%	-0.18%	-0.18%	-0.20%	-0.18%	-0.25%	0.04%	-0.56%	22
5	-0.36%	-0.26%	-0.25%	-0.28%	-0.26%	-0.38%	0.13%	-0.90%	18
6	-0.28%	-0.30%	-0.30%	-0.35%	-0.30%	-0.50%	0.30%	-1.33%	14
7	-0.60%	-0.24%	-0.21%	-0.42%	-0.24%	-0.80%	1.46%	-3.23%	10
8	-0.47%	-0.19%	-0.14%	-0.51%	-0.19%	-1.00%	2.19%	-4.49%	6
Menor	0.67%	0.84%	0.85%	0.77%	0.84%	0.52%	1.68%	-0.69%	8

Com relação à comparação do modelo produzido pelo processo com outros algoritmos de redução de características, para o Contexto C1, o modelo produzido pelo processo demonstrou resultados ligeiramente superiores a todos os outros algoritmos na acurácia e F1. No entanto, para a métrica F1 da classe evasão, o resultado foi inferior ao modelo produzido com as características do algoritmo ReliefF, conforme ilustrado na Figura 7.

O último contexto (C2) é um contexto especializado dentro do campo dos cursos de enfermagem, com foco exclusivo em instituições de ensino superior públicas, com o escopo geográfico limitado ao estado de São Paulo. A Tabela 11 resume os resultados do

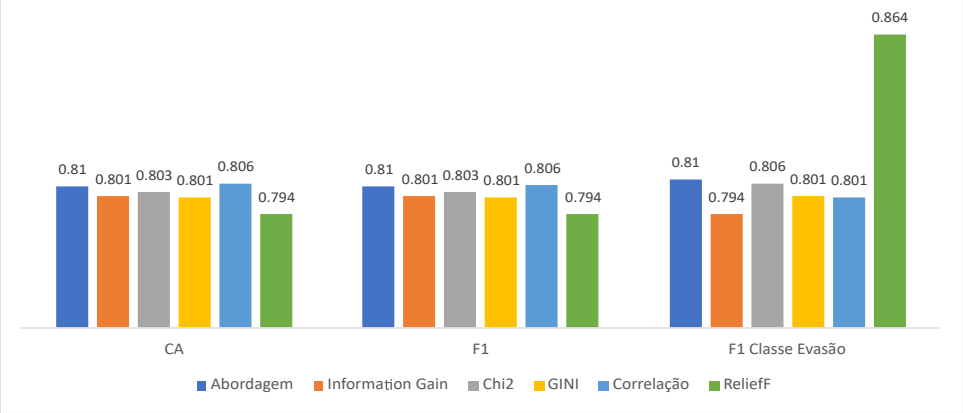

Figura 7 – Comparação entre métricas produzidas pelo melhor modelo para o Contexto C1 com outros algoritmos de seleção de atributos

experimento. Neste contexto, ao contrário dos outros, o menor modelo consistentemente alcançou resultados superiores em todas as métricas em comparação com o modelo inicial. Notavelmente, ele demonstrou o maior *recall* em comparação com os outros contextos, e a acurácia do melhor modelo superou a do modelo inicial em mais de 2%, usando apenas 15% das características iniciais. É importante observar que o menor modelo e a sexta iteração têm as mesmas características. No entanto, a tarefa de classificação foi executada novamente, resultando em novos resultados com valores de métricas muito próximos.

Tabela 11 – A diferença (%) da performance entre o melhor modelo em cada iteração e o modelo inicial para o contexto de cursos públicos de enfermagem. A coluna QC indica a quantidade de características.

	Média total das classes					**Classe Evasão**			
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall	**QC**
Inicial	0.911	0.840	0.840	0.840	0.840	0.841	0.834	0.849	44
1 (boruta)	-0.82%	-0.85%	-0.85%	-0.84%	-0.85%	-0.61%	-1.90%	0.67%	27
2	-1.29%	-1.19%	-1.19%	-1.19%	-1.19%	-1.15%	-1.29%	-1.01%	23
3	-1.52%	-2.55%	-2.55%	-2.55%	-2.55%	-2.27%	-3.88%	-0.67%	19
4	-2.31%	-2.89%	-2.89%	-2.88%	-2.89%	-2.70%	-3.72%	-1.68%	15
5	-1.23%	-1.19%	-1.19%	-1.21%	-1.19%	-0.79%	-2.99%	1.35%	11
6	-1.47%	-2.04%	-2.04%	-2.04%	-2.04%	-2.02%	-2.02%	-2.02%	7
7	2.59%	-0.85%	-0.80%	-1.36%	-0.85%	0.38%	-6.76%	6.73%	3
Menor	-1.37%	-2.04%	-2.04%	-2.08%	-2.04%	-2.16%	-1.32%	-2.68%	7

Com relação à comparação do modelo produzido pelo processo com outros algoritmos de redução de características, para o Contexto C2, o modelo produzido pelo processo demonstrou resultados superiores a todos os outros algoritmos em todas as métricas comparadas, como ilustrado na Figura 8.

Por fim, um último resultado apresentado na Figura 9 fornece uma visão geral

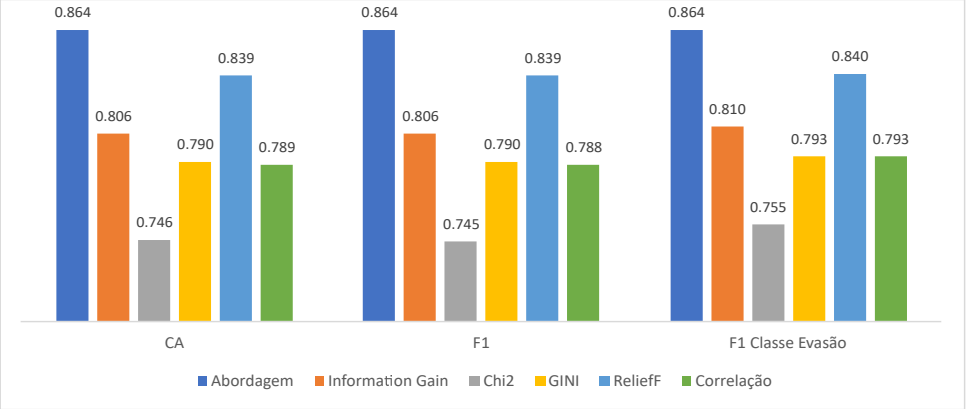

Figura 8 – Comparação entre métricas produzidas pelo melhor modelo para o Contexto C2 com outros algoritmos de redução de características

dos resultados em todos os contextos, mostrando a acurácia produzida pelo modelo inicial, juntamente com o resultado produzido pelo modelo gerado com as características consideradas relevantes pelo algoritmo Boruta, e os resultados dos melhores e menores modelos identificados para cada contexto específico, incluindo o número de características utilizadas em cada modelo.

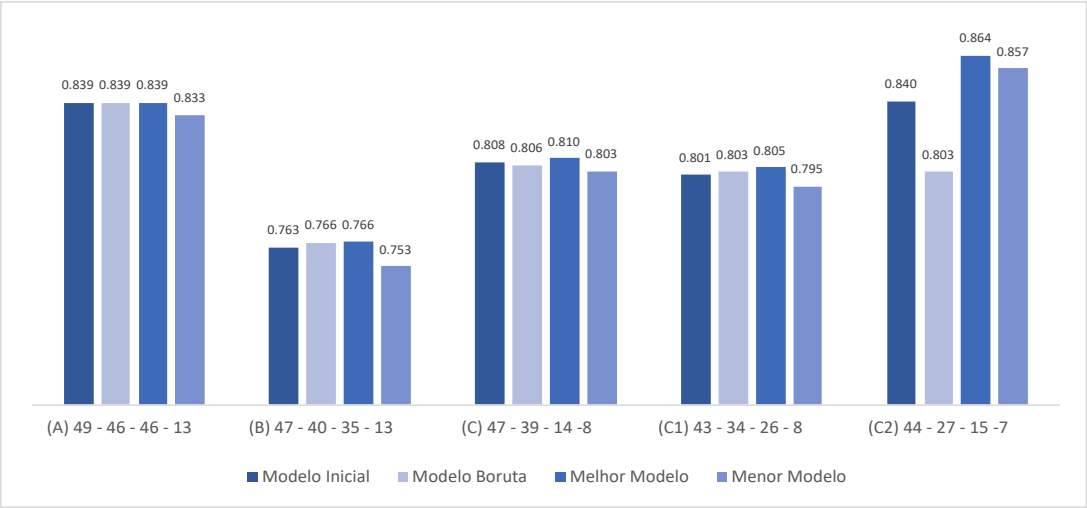

Figura 9 – Acurácia dos modelos iniciais, modelos com as características do Boruta, melhores e menores modelos em diferentes contextos

Em todos os contextos, os modelos que apresentaram melhores resultados tiveram um número reduzido de características em comparação com o modelo inicial e o modelo construído com as características definidas como relevantes pelo Boruta, exceto pelo contexto A, no qual o melhor modelo foi construído com as características do Boruta. Esse resultado é um indicativo significativo de que o processo proposto é eficaz, uma vez que melhorou os resultados produzidos pelo algoritmo de seleção de atributos.

Isso sugere que a metodologia desenvolvida foi capaz de aprimorar a seleção de

atributos, resultando em modelos de melhor desempenho. Ainda com relação aos resultados apresentados na 9, percebe-se que o no contexto C2, o menor modelo apresenta resultados melhores que o modelo inicial e que nos contextos mais amplos os melhores modelos possuem um número maior de características.

5.1 Discussão sobre o modelo proposto

Os resultados retratados na Figura 9 revelam que, em todos os contextos, o modelo ótimo utilizou um número reduzido de características em comparação com o modelo inicial.

Para o contexto A, o modelo ótimo removeu 3 características, mantendo a acurácia. Para o contexto B, o modelo ótimo removeu 12 características, tendo uma acurácia superior em 0.3 pontos percentuais (PP), enquanto o menor modelo teve a acurácia inferior em 1 PP, porém com 34 características a menos, sendo ainda um modelo aceitável.

Já nos contextos C e C1, o melhor modelo foi superior, respectivamente, em 0.2 e 0.4 PP, com 33 e 17 menos características, enquanto o menor modelo teve uma acurácia inferior em 0.5 e 0.6 PP, com 39 e 35 menos atributos.

Observando o contexto C2, que é o contexto mais específico dos contextos elencados neste trabalho, o modelo ótimo (15 características) foi superior em 2.4 PP que o modelo inicial (44 características) e 6.1 PP superior ao Boruta (27 características). O menor modelo, com apenas 7 características, foi superior ao modelo inicial e ao Boruta em 1.7 e 5.4 PP.

Ainda de acordo com a Figura 9, em quatro dos cinco contextos, o modelo com melhor desempenho empregou menos características do que o Boruta. Pela ótica da utilização apenas do Boruta, observa-se que, no contexto C, o Boruta foi inferior ao modelo inicial, podendo dar uma falsa sensação de que não há como reduzir as características do modelo. Contudo, o melhor modelo do processo proposto reduziu consideravelmente as característica e ainda teve uma acurácia maior que o modelo inicial e o Boruta, sugerindo resultados promissores para a redução de características.

Assim, em geral o processo conseguiu gerar modelos com mais acurácia que os modelos iniciais e que os modelos gerados com a remoção das características indicadas como não importantes pelo algoritmo Boruta. Pelos experimentos produzidos pode-se então considerar que o processo apresenta uma melhoria ao algoritmo de redução de atributos Boruta.

Como forma de verificar a eficácia do processo proposto, além da comparação direta com os modelos iniciais e os modelos gerados pelo Boruta, também comparamos os melhores modelos de cada contexto com o resultado de outros cinco algoritmo de seleção

de atributos, conforme mostram as Figuras 4, 5, 6, 7, e 8. Em geral, percebe-se que o processo proporcionou resultados similares ou melhores que os algoritmos comparados. Vale ressaltar que essa comparação foi feita utilizando a mesma quantidade de características do melhor modelo, como descrito na Seção Seção 5 e todos os modelos foram treinados e testados com o mesmo processo, utilizando do algoritmo *Random Forest*.

Para os contextos A e B, que são contextos mais abrangentes, os resultados para todos os algoritmos foram muito próximos, sem uma diferença significativa entre eles. Já em contextos mais específicos houve uma mudança de comportamento. No contexto C, o processo proposto foi 4 PP superior em relação ao segundo melhor algoritmo (ReliefF) e 11 PP melhor que o pior algoritmo (Chi2) em todas as métricas.

Já no contexto C1, o processo foi ligeiramente superior nas métricas acurácia e F1, porém o ReliefF foi 6 PP superior na métrica F1 para a classe de evasão. Observando o resultado do contexto C2, o processo foi 2.5 PP superior em relação ao segundo melhor algoritmo (ReliefF) para todas as métricas e 12 PP melhor em relação ao pior algoritmo (Chi2).

Com isso, observa-se que os melhores modelos gerados pelo processo desempenharam melhor em contextos mais específicos, tendo uma diferença significativa nos contextos C e C2.

Além disso, conforme representado na Figura 10, três métricas são apresentadas visualmente, mostrando seu comportamento em relação ao número de características em dois contextos. No Contexto C2, os resultados exibem uma melhoria à medida que o número de características diminui, atingindo seu pico em 15 características. No entanto, além desse limite, a acurácia começa a declinar, embora experimente uma leve melhoria com 7 características.

Por outro lado, no Contexto C, a acurácia dos modelos permanece relativamente estável até atingir 14 características, onde alcança seu desempenho ótimo. Dessa forma, por meio da Figura 10, percebe-se que não existe uma padrão claro entre a diminuição da características e a acurácia, o que em parte explica os resultados conseguidos pelo processo se comparado com outros algoritmos de redução de características.

Em relação aos resultados, é relevante enfatizar a importância do processo proposto na criação de modelos de alta precisão por meio da redução de características, pois os algoritmos de seleção de atributos nem sempre garante a geração dos melhores modelos. O processo proposto produziu resultados superiores em comparação com o uso isolado de algoritmos de seleção de atributos.

Dado que o processo proposto neste trabalho gerou modelos menores que os modelos iniciais e originados pelo Boruta na maioria dos contextos, sendo equivalentes ou superiores; e dado que na comparação entre algoritmos de seleção de atributos com o

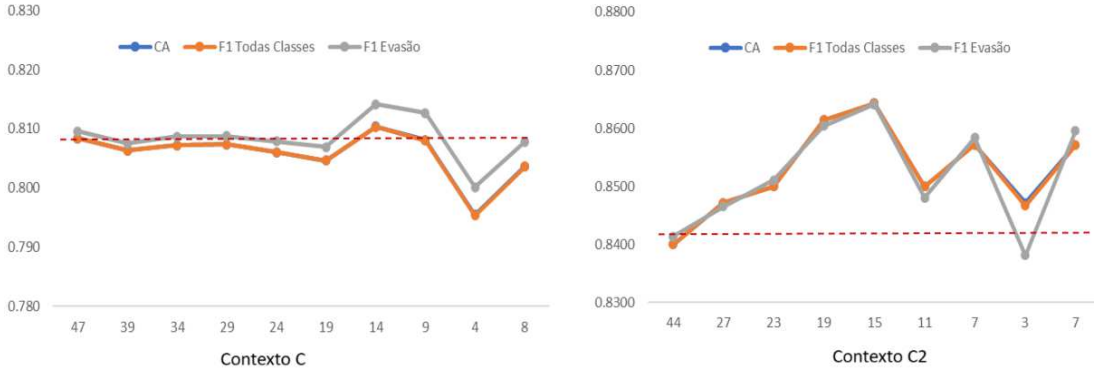

Figura 10 – Resultados das métricas para os contextos C e C2 em relação ao número de características

As Figuras mostram os resultados de CA, F1 sobre todas as classes e F1 sobre a classe de evasão considerando as iterações do processo. A linha vermelha tracejada indica o resultado do modelo inicial e o eixo x o número de características.

processo proposto, o modelo proposto, em geral, foi similar ou superior, fica então evidenciada a contribuição computacional para reduzir dimensões em um conjunto de dados mantendo uma acurácia similar ou superior de modelos inciais.

5.2 Características Importantes para a Predição de Evasão

Neste estudo, implementou-se um processo com o objetivo de minimizar o número de características, ao mesmo tempo em que foi mantida a acurácia do modelo comparável à do modelo original contendo todas as características. Esse processo foi aplicado em cinco contextos distintos, e em nenhum deles os modelos resultantes apresentaram uma diminuição na acurácia de mais de 1.34% em comparação com o modelo inicial. Essa metodologia nos permitiu desenvolver modelos preditivos utilizando apenas as características mais pertinentes, essenciais para a previsão de evasão.

Quanto às características mais importantes para prever a evasão, em primeiro lugar, como mostrado na Figura 11, há uma diferença no número de características nos menores modelos e melhores entre os contextos. Assim, isso é uma clara indicação de que existem diferenças na importância das características em contextos distintos. Além disso, nos contextos mais heterogêneos, nos melhores e menores modelos, há um maior número de características, indicando que quanto mais amplo o contexto, mais difícil é prever os fatores que levam à evasão.

Dada a natureza variável do número de características nos modelos menores em diferentes contextos, uma comparação direta torna-se inviável. Portanto, a Figura 12 mostra a matriz de correlação, para representar a relação entre as características identificadas nos menores modelos de cada contexto. Assim, diferentes contextos apresentam pesos distintos nas características para a previsão de evasão. Mesmo contextos semelhantes apresentam

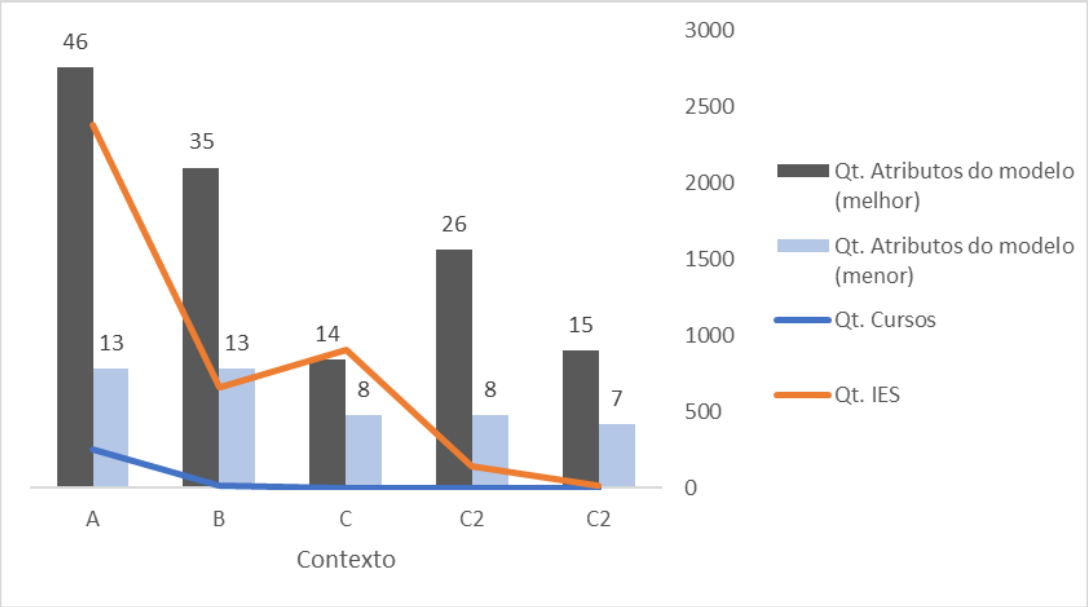

Figura 11 – Resultado das métricas para cada contexto em relação ao número de características

diferenças significativas nas características mais importantes.

Contextos	A	B	C	C1	C2
A	1.00	0.69	0.87	0.75	0.71
B	0.69	1.00	1.00	0.88	0.86
C	0.53	0.62	1.00	0.75	0.86
C1	0.46	0.54	0.75	1.00	0.57
C2	0.38	0.46	0.38	0.50	1.00

Figura 12 – Matriz de correlação das características mais importantes em diferentes contextos

A Tabela 12 fornece uma visão geral das características mais importantes considerando os cinco contextos explorados, além da origem de cada característica (se é relativa ao aluno, curso ou IES). De acordo com as descobertas do estudo, a correlação entre a duração da matrícula de um aluno e a porcentagem do curso concluída emerge como a característica mais importante para prever a evasão. Além disso, as atividades extracurriculares são essenciais para prever a evasão, embora em menor grau em contextos onde apenas IES privadas estejam envolvidas. A idade do aluno também aparece como uma característica importante, assim como a duração do curso.

As características relacionadas às IES também são altamente significativas. Isso inclui o código da IES e a in_busca_integrada, uma característica que indica o nível de infraestrutura da IES. Além disso, nos contextos onde a análise foi realizada para cursos

do Brasil todo, características relacionadas à localidade também são fundamentais.

Nos contextos mais homogêneos, observa-se que características vinculadas aos estudantes desempenham um papel mais significativo e contribuem para a criação de modelos com mais acurácia. Essas características dos estudantes podem ser divididas em duas categorias: aquelas relacionadas ao seu progresso no curso (tempo_entrada_curso, porc_concluida_range, in_atividade_extracurricular, in_concluinte, ingresso_processo, tp_turno) e características pessoais (idade, tp_cor_raca). Os resultados indicam que tp_cor_raca é importante em contexto privado, enquanto tp_turno no contexto público.

A discussão final diz respeito às descobertas de outros estudos sobre evasão e as comparações com este estudo. A Tabela 2, da Seção 2.4, resume os dados de vários estudos que se concentraram na previsão de evasão no ensino superior presencial, destacando métricas como AUC, acurácia, preicssão e *recall*, e as características-chave de cada estudo.

Neste estudo, as métricas apresentaram os seguintes valores: AUC variando de 0,85 a 0,93; CA de 0,77 a 0,86; F1 de 0,77 a 0,86; Acurácia de 0,77 a 0,86; e *Recall* de 0,77 a 0,86. Ao comparar os resultados deste estudo com outros encontrados na literatura, as métricas obtidas em neste estudo são equivalentes ou superiores às de outros estudos. No entanto, é importante reconhecer que muitos estudos têm contextos que diferem significativamente dos contextos deste estudo, incluindo variações nos dados utilizados.

Vários estudos incorporam dados financeiros e de desempenho acadêmico, aspectos que não foram abordados neste estudo. O único trabalho que utiliza uma fonte de dados similar ao utilizada neste estudo é apresentado por Teodoro *et al.* (2020) [30]. Contudo, em quatro dos cinco contextos analisados neste estudo, os resultados foram superiores ao conseguidos em [30].

Em relação às características mais significativas, conforme indicado pelos estudos na Tabela 2, fatores como desempenho acadêmico e, no caso de IES privadas, dados financeiros, podem contribuir para aprimorar a previsão de evasão, juntamente com as características consideradas importantes neste trabalho.

Observa-se que, de acordo com o conjunto de dados explorado, as principais características relacionada à evasão de aluno são: tempo_entrada_curso e porc_concluida_range. Ambas estão presentes nos modelos de todos os contextos e são atributos de origem Aluno, relacionadas com o tempo que o aluno está no curso e qual é a porcentagem do curso concluída. Ou seja, são características individuais do aluno, porém diretamente relacionadas com a situação temporal e de conclusão do curso.

Ainda sobre características importantes na maioria dos modelos, algumas características estão presentes em quatro dos cinco contextos, tendo os contextos A e B em todos os cenários, com uma variação entre os contextos C, C1 e C2. As características são: in_atividade_extracurricular, in_ingresso_processo, in_concluinte, co_uf e idade.

Tabela 12 – Características mais importantes e os contextos em que apareceram no menor modelo

Característica	**Descrição**	**Contextos**	**Origem**
tempo_entrada_-curso	Tempo total em anos que o aluno ingressou no curso	A;B;C;C1;C2	Aluno
porc_concluida_-range	Descreve a porcentagem total concluída pelo aluno do curso em faixas.	A;B;C;C1;C2	Aluno
in_atividade_extracurricular	Informa se o aluno participa de algum tipo de atividade extracurricular (estágio não obrigatório, extensão, monitoria e pesquisa)	A;B;C;C2	Aluno
in_ingresso_processo	Determina se entrou no curso por meio de algum processo seletivo ou não	A;B;C;C1	Aluno
in_concluinte	Informa se o aluno é concluinte	A;B;C1;C2	Aluno
co_uf	Codigo do IBGE da unidade da federação do local de oferta do curso presencial	A;B;C;C1	Curso/IES
idade	faixa etaria do aluno	A;B;C1;C2	Aluno
in_busca_integrada	Informa se as bibliotecas da IES oferecem serviços pela internet	A;B;C	IES
co_ies	Código único de identificação da IES	A;B;C	IES
tempo_curso_-range	Faixa de tempo que o curso está em funcionamento	B;C;C1	Curso
in_ingresso_total	Informa se o aluno é ingressante no curso, não importando a forma de ingresso utilizada.	B;C2	Aluno
cargar_horaria_-curso	Carga horária total do cuso categorizada em faixas	A	Curso
in_repositorio_-institucional	Informa se a IES possui base de dados online que reúne de maneira organizada a produção científica da instituição	A	IES
co_cine_rotulo	Código de identificação do curso, conforme adaptação da Classificação Internacional Normalizada da Educação Cine/Unesco	A	Curso
co_municipio	Código do IBGE do municipio do local de oferta do curso presencial	A	Curso/IES
no_curso	Nome do curso	B	Curso
co_regiao	Codigo do IBGE da região da federação do local de oferta do curso presencial	B	Curso/IES
tp_turno	Tipo do turno do curso ao qual o aluno está vinculado	C2	Curso
tp_cor_raca	Tipo da cor/raça do aluno	C1	Aluno

Dessas cinco características, apena a co_uf tem como origem Curso/IES, indicando a unidade federativa de oferta do curso. A característica idade se destaca neste conjunto, pois é intrínseca ao aluno, não tendo relação com Curso ou IES.

Presentes em três contextos, as características in_busca_integrada, co_ies e tempo_curso_range são observadas. Todas aparecem nos contextos B e C, havendo uma variação entre os contextos A e C1. Neste caso, são características diretamente relacionadas à IES ou Curso, não tendo uma associação direta com o Aluno.

A característica in_ingresso_total aparece em dois contextos, enquanto o restante das características da Tabela 12 aparece em um contexto apenas. Dessas características, apenas uma (tp_cor_raca) é intrínseca ao aluno (juntamente com a idade, mais difundida nos contextos). As características restantes são intrínsecas ao Curso ou IES.

Em resumo, características relacionadas aos alunos aparecem mais frequentemente nos contextos, o que permite entender que o fenômeno da evasão pode ter características comuns em diversos contextos, enquanto características de Curso e IES podem sofrer uma maior variação a depender do contexto em que se queria analisar a evasão.

Realizando uma comparação com trabalhos relacionados presentes na Tabela 2, observa-se algumas características importantes neste trabalho também foram identificadas em outros trabalhos. São elas: etnia [45], ano de acesso [47], atividade extracurricular [30], idade [30] e carga horária do curso [30]. Observa-se que em [30] existem três características comuns, pois foi utilizado a mesma fonte de dados que este trabalho.

É importante destacar que este trabalho apresenta duas diferenças-chave em comparação com os estudos encontrados na literatura. Em primeiro lugar, ele introduz um processo que permite a redução do número de características para o mínimo necessário, enquanto ainda produz modelos com acurácia semelhante ou superior a outros trabalhos. Em segundo lugar, ele é o único estudo a analisar a evasão em diversos contextos, indicando que as razões que levam à evasão variam de acordo com o cenário de análise.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O principal objetivo deste trabalho foi propor um processo iterativo para gerar modelos preditivos com dimensões reduzidas, que priorizam as características mais relevantes do conjunto de dados. O processo foi aplicado em um conjunto de dados do ensino superior, tendo como alvo a predição da evasão para poder validar o processo.

Considerandos os experimentos realizados e os resultados apresentados no decorrer do trabalho, o processo indica que produz resultados superiores ao algoritmo Boruta, o qual o processo se baseia. Além disso, também apresenta melhores resultados que outros cinco algoritmos de seleção de atributos comparados.

Com relação à análise da evasão, que foi o foco de aplicação do processo, os resultados sugerem que em contextos mais restritos, modelos com menos características tendem a produzir previsões com mais acurácia. Tipicamente, esses modelos restritos são baseados em dados do aluno. Por outro lado, em contextos heterogêneos, os modelos utilizam mais dados do curso e da instituição de ensino superior. A importância da característica está relacionada a vários fatores, como cobertura, área, curso, tipo de instituição e modalidade de ensino, entre outros.

Dessa forma, considerando os resultados apresentados, este trabalho contribui de várias maneiras. Em primeiro lugar, introduz um processo centrado na seleção de características. Nos experimentos realizados, este processo produziu modelos reduzidos com maior acurácia do que os modelos originais e aqueles gerados apenas pela remoção de características não importantes identificadas pelos algoritmos de seleção de características.

Em contraste com estudos relacionados anteriores, esta pesquisa explora a evasão em diversos cenários, que variam de contextos amplos e heterogêneos a contextos restritos e homogêneos. Isso nos permite oferecer evidências destacando a dependência contextual da evasão e a necessidade de características específicas em cada contexto para a predição da evasão. No entanto, apesar dessa dependência de contexto, este estudo indica que, para cursos presenciais no Brasil, existe um conjunto central de características críticas para prever as taxas de evasão em todos os contextos.

Embora criar modelos preditivos para evasão seja indiscutivelmente importante, também é crucial compreender os principais fatores associados a ela para tomar decisões informadas com o objetivo de prevenir a evasão. Nesse contexto, este estudo oferece um processo que não apenas auxilia na previsão da evasão, mas também oferece aos tomadores de decisão um conjunto conciso de características críticas que a influenciam. Esse benefício duplo de poder preditivo e seleção de características simplificada aprimora a utilidade prática desta pesquisa.

Como principais trabalhos futuros para esta pesquisa, vislumbra-se aplicar o processo em outros contextos para ter uma compreensão mais abrangente da predição de evasão, além de aplicar também em conjuntos de dados de diferentes domínios. Outro possível trabalho futuro é incorporar dados de desempenho acadêmico ao *dataset*, pois conforme apresentado nos trabalhos relacionados essas características tem potencial para melhorar a acurácia dos modelos produzidos.

Por fim, em relação ao processo proposto, há a possibilidade de automatizá-lo, encapsulando-o (Figura 3) em funções *R* ou *Python*. Com isso, passos manuais serão eliminados, proporcionando uma maior velocidade na execução do processo, além de torná-lo acessível aos pares.

BIBLIOGRAFIA

[1] VAARMA, M.; LI, H. Predicting student dropouts with machine learning: An empirical study in finnish higher education. *Technology in Society*, v. 76, p. 102474, 2024. ISSN 0160-791X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0160791X24000228>.

[2] CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 1, p. 16–28, 2014.

[3] GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003.

[4] MENOLLI, A. et al. Bi–based methodology for analyzing higher education: A case study of dropout phenomenon in information systems courses. In: *XVI Brazilian Symposium on Information Systems*. [S.l.: s.n.], 2020. p. 1–8.

[5] DEMETER, E. et al. Predicting first-time-in-college students' degree completion outcomes. *Higher Education*, Springer, p. 1–21, 2022.

[6] MUSSO, M. F.; HERNÁNDEZ, C. F. R.; CASCALLAR, E. C. Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education*, Springer, v. 80, p. 875–894, 2020.

[7] PEREZ, B.; CASTELLANOS, C.; CORREAL, D. Applying data mining techniques to predict student dropout: A case study. In: *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. [S.l.: s.n.], 2018. p. 1–6.

[8] INEP. *Instruções para utilização dos Microdados do Censo da Educação Superior*. [S.l.], 2019. 1–9 p.

[9] FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, p. 641–659, 2007.

[10] WANG, Y. Big opportunities and big concerns of big data in education. *TechTrends*, Springer, v. 60, p. 381–384, 2016.

[11] INEP. *Microdados do Censo da Educação Superior 2019 - Manual do Usuário*. [S.l.], 2019.

[12] RODRÍGUEZ-MUÑIZ, L. J. et al. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *Plos one*, Public Library of Science San Francisco, CA USA, v. 14, n. 6, p. e0218796, 2019.

[13] VILORIA, A. et al. Integration of data technology for analyzing university dropout. *Procedia Computer Science*, Elsevier, v. 155, p. 569–574, 2019.

[14] CASTRO, T. R. Metodologia de acompanhamento e combate à evasão: O caso do curso de engenharia de produção da unespar. *Revista de Ensino de Engenharia*, v. 40, 2021.

[15] CARVALHO, A. F.; LOPES, F. M.; SOUTO, M. C. A machine learning approach to predict dropout in distance education. *Computers & Education*, Elsevier, v. 132, p. 49–64, 2019.

[16] BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

[17] LEE, H.-C.; WU, Y.-T.; WU, T.-Y. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, Elsevier, v. 58, n. 1, p. 176–186, 2012.

[18] ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor & Francis, v. 46, n. 3, p. 175–185, 1992.

[19] ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 40, n. 6, p. 601–618, 2010.

[20] CORTES, C.; VAPNIK, V. N. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

[21] RISH, I. An empirical study of the naive bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, v. 3, p. 41–46, 2001.

[22] MOHANTY, M. R.; MISHRA, A. K. Prediction of student's performance using naive bayes classifier. *International Journal of Computer Applications*, Foundation of Computer Science (FCS), v. 139, n. 4, 2016.

[23] BISONG, E.; BISONG, E. Logistic regression. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, Springer, p. 243–250, 2019.

[24] AMORIM, P. et al. Mining student data to predict dropout-prone courses in higher education institutions: A decision tree-based approach. In: IEEE. *2016 IEEE Frontiers in Education Conference (FIE)*. [S.l.], 2016. p. 1–5.

[25] KURSA, M. B.; RUDNICKI, W. R. Feature selection with the boruta package. *Journal of statistical software*, v. 36, p. 1–13, 2010.

[26] JOVIć, A.; BRKIć, K.; BOGUNOVIć, N. A review of feature selection methods with applications. In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. [S.l.: s.n.], 2015. p. 1200–1205.

[27] ANAND, N. et al. Feature selection on educational data using boruta algorithm. *International Journal of Computational Intelligence Studies*, Inderscience Publishers (IEL), v. 10, n. 1, p. 27–35, 2021.

[28] BHALAJI, N.; KUMAR, K. S.; SELVARAJ, C. Empirical study of feature selection methods over classification algorithms. *International Journal of Intelligent Systems Technologies and Applications*, Inderscience Publishers (IEL), v. 17, n. 1-2, p. 98–108, 2018.

[29] FRANCO, J. J. et al. Usando mineraçao de dados para identificar fatores mais importantes do enem dos últimos 22 anos. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 1112–1121.

[30] TEODORO, L. 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, 2020. Disponível em: <http://dx.doi.org/10.5753/rbie.2020.28.0.838>.

[31] MARTINS, D. A.; LEITE, L. P.; LACERDA, C. B. F. d. Políticas públicas para acesso de pessoas com deficiência ao ensino superior brasileiro: uma análise de indicadores educacionais. *Ensaio: avaliação e políticas públicas em educação*, SciELO Brasil, v. 23, p. 984–1014, 2015.

[32] LASSIBILLE, G.; GÓMEZ, L. N. Why do higher education students drop out? evidence from spain. *Education Economics*, Taylor & Francis, v. 16, n. 1, p. 89–105, 2008.

[33] RUMBERGER, R. W. The economics of high school dropouts. *The economics of education*, Elsevier, p. 149–158, 2020.

[34] COSTA, F. J. d.; BISPO, M. d. S.; PEREIRA, R. d. C. d. F. Dropout and retention of undergraduate students in management: a study at a brazilian federal university. *RAUSP Management Journal*, SciELO Brasil, v. 53, p. 74–85, 2018.

[35] KEHM, B. M.; LARSEN, M. R.; SOMMERSEL, H. B. Student dropout from universities in europe: A review of empirical literature. *Hungarian Educational Research Journal*, Akadémiai Kiadó Budapest, v. 9, n. 2, p. 147–164, 2019.

[36] LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, v. 25, p. 14, 2012.

[37] CALIXTO, C. Análise das causas de evasão discente no curso de licenciatura em computação: um estudo da ufpb virtual no formato uab. *Revista Tecnologias na Educação*, v. 7, n. 12, p. 1–13, 2015.

[38] MENOLLI, A.; NETO, J. C. Uma análise do perfil dos cursos de licenciatura em computação no brasil. *Revista Brasileira de Informática na Educação*, v. 29, p. 01–24, 2021.

[39] OLIVEIRA, J. L. et al. Undergraduate students' effectiveness in an institution with high dropout index. In: IEEE. *2020 IEEE Frontiers in Education Conference (FIE)*. [S.l.], 2020. p. 1–7.

[40] BERKA, A.; MAREK, M. 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, 2021. Disponível em: <http://dx.doi.org/10.5753/rbie.2020.28.0.838>.

[41] CANNISTRà, T. C.; SILVA, J. C.; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Computação Aplicada*, v. 10, n. 3, p. 11–20, 2018.

[42] DELEN, D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, Elsevier, v. 49, n. 4, p. 498–506, 2010.

[43] DJULOVIC, A.; LI, D. Towards freshman retention prediction: a comparative study. *International Journal of Information and Education Technology*, IACSIT Press, v. 3, n. 5, p. 494–500, 2013.

[44] MARTINS, M. V. et al. Multi-class phased prediction of academic performance and dropout in higher education. *Applied Sciences*, MDPI, v. 13, n. 8, p. 4702, 2023.

[45] MATZ, S. C. et al. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, Nature Publishing Group UK London, v. 13, n. 1, p. 5705, 2023.

[46] NAGY, M.; MOLONTAY, R. Interpretable dropout prediction: Towards xai-based personalized intervention. *International Journal of Artificial Intelligence in Education*, Springer, p. 1–27, 2023.

[47] SONG, Z. et al. All-year dropout prediction modeling and analysis for university students. *Applied Sciences*, MDPI, v. 13, n. 2, p. 1143, 2023.

[48] YU, R.; LEE, H.; KIZILCEC, R. F. Should college dropout prediction models include protected attributes? In: *Proceedings of the eighth ACM conference on learning@ scale*. [S.l.: s.n.], 2021. p. 91–100.

[49] RSTUDIO, P. *RStudio: Open source and enterprise-ready professional software for the R community*. 2023. Disponível em: <https://www.rstudio.com/>.

[50] POSTGRESQL. *PostgreSQL: The world's most advanced open source database*. 2023. Disponível em: <https://www.postgresql.org/>.

[51] HITACHIVANTARA. *Pentaho Data Integration*. 2017. Disponível em: <https://help.hitachivantara.com/Documentation/Pentaho/7.1/0D0/Pentaho_Data_Integration>.

[52] Python Software Foundation. *Python*. 2023. Acessado em 30 de outubro de 2023. Disponível em: <https://www.python.org>.

[53] VORA, S.; YANG, H. A comprehensive study of eleven feature selection algorithms and their impact on text classification. In: IEEE. *2017 Computing Conference*. [S.l.], 2017. p. 440–449.