MESTRADO EM SISTEMAS E COMPUTAÇÃO

CELSO BARRETO DA SILVA

UM MÉTODO DE REFERÊNCIA COM APRENDIZADO DE MÁQUINA PARA PREVISÃO DE EVASÃO DE ALUNOS EM INSTITUIÇÕES DE ENSINO SUPERIOR PARTICULAR

Salvador
2023

CELSO BARRETO DA SILVA

UM MÉTODO DE REFERÊNCIA COM APRENDIZADO DE MÁQUINA PARA PREVISÃO DE EVASÃO DE ALUNOS EM INSTITUIÇÕES DE ENSINO SUPERIOR PARTICULAR

Dissertação apresentada ao Programa de Pós-Graduação em Sistemas e Computação da Universidade Salvador - UNIFACS, como requisito parcial à obtenção do título de Mestre.

Orientador: Prof. Dr. Joberto Sérgio Barbosa Martins.

Coorientadores: Prof. Dra. Ana Patrícia Fontes Magalhães Mascarenhas.

Prof. Dr. José Vicente Cardoso Santos.

Salvador
2023

FICHA CATALOGRÁFICA
(Elaborada pelo Sistema de Bibliotecas da UNIFACS Universidade Salvador,

Silva, Celso Barreto da

Um método de referência com aprendizado de máquina para previsão de evasão de alunos em instituições de ensino superior particular/ Celso Barreto da Silva. – Salvador: Unifacs, 2023.

134 f.: il.

Dissertação apresentada ao Programa de Pós-Graduação em Sistemas e Computação da Universidade Salvador - UNIFACS, como requisito parcial para obtenção do grau de Mestre.

Orientador: Prof. Dr. Joberto Sérgio Barbosa Martins.
Coorientadores: Prof. Dra. Ana Patrícia Fontes Magalhães Mascarenhas.
Prof. Dr. José Vicente Cardoso Santos.

1. Educação superior - evasão. 2. Aprendizado de máquina. I. Martins, Joberto Sérgio Barbosa, orient. II. Mascarenhas, Ana Patrícia Fontes Magalhães, co-orient. III. Santos, José Vicente Cardoso, co-orient. IV. Título.

CDD: 388

CELSO BARRETO DA SILVA

UM MÉTODO DE REFERÊNCIA COM APRENDIZADO DE MÁQUINA PARA PREVISÃO DE EVASÃO DE ALUNOS EM INSTITUIÇÕES DE ENSINO SUPERIOR PARTICULAR

Dissertação apresentada ao Programa de Pós-Graduação em Sistemas e Computação da Universidade Salvador - UNIFACS, como requisito parcial para obtenção do grau de Mestre aprovada pela seguinte banca examinadora:

Joberto Sérgio Barbosa Martins _____
Doutor em Ciência da Computação pela Université Paris VI – UPMC
Universidade Salvador - UNIFACS

Ana Patrícia Fontes Magalhães Mascarenhas _____
Doutora em Ciências da Computação pela Universidade Federal da Bahia (UFBA)
Universidade do Estado da Bahia (UNEB)

Artur Henrique Kronbauer _____
Doutor em Ciência da Computação pela Universidade Federal da Bahia (UFBA)
Universidade Salvador - UNIFACS

Rafael Freitas Reale _____
Doutor em Ciência da Computação pela Universidade Federal da Bahia (UFBA)
Instituto Federal de Educação da Bahia (Valença)

Salvador, de de 2023.

Dedico a Deus meu criador, minha inspiração que mesmo diante das adversidades me sustentou e colocou em meu coração a vontade de vencer e tornar possível cada sonho outrora impossível.

Com eterna gratidão e amor dedico em especial a minha esposa, ombro amigo, voz consoladora, coração generoso, minha princesinha e meu maior presente de Deus e com seu sorriso diário tem me ajudado a suportar as duras batalhas.

As minhas filhas Michelle e Raquel, minhas pequenas, meus preciosos tesouros que sempre me dão motivos não desistir, nunca!

Ao meu amado e inesquecível pai (*in memoriam*), seus ensinos eu levo adiante conquistando o impossível, espero que ouça que o seu trabalho, valeu a pena!

As mães Marias (Edina e José), com tarefas distintas em minha vida, geração e vida e criação. Sem vocês, eu não estaria aqui. Obrigado Mães!

Ao meu avô (Brazilino), amei desde o início que conheci e minha vovó (Tomázia). Obrigado.

A minha vovó amada paterna Marculina Azevedo, eu nunca a esquecerei enquanto vida eu tiver. De onde esteja, saiba que ainda sinto o seu abraço e carinho vó. Te amo!

As cunhadas Simone e Célia, que sempre com sorrisos e muita alegria me ajudaram mesmo sem saber, Obrigado!

Meus irmãos Jeferson (Joãozinho), Alberto (Gão), Allan (Chang) e Manuel (Nel) que sempre serviram de apoio em muitos momentos de cansaço, serviram de inspiração, serviram de modelo, serviram de apoio, aos meus manos orelhas (orêas) inesquecíveis além da gratidão eu digo - **“BORA!”.**

AGRADECIMENTOS

Ao ETERNO criador por permitir realizar esse sonho e subir este degrau em minha vida;

Aos familiares, mães, amado pai (*in memoriam*), amados irmãos, sobrinhos, tios, tias, primos, cunhados, amada esposa;

A minha cunhada Simone Lima e meu irmão Manuel, se existir alguma palavra acima de obrigado, usarei neste momento.

Aos amigos de faculdade, em especial, pelos momentos de alegrias, dificuldade e grandes aprendizados;

Aos amigos(as) pessoais, em especial Sidnei Barros, Cevaldo Santos e Santos e Marcos Leite;

A educadora Ana Patrícia Fontes Magalhães Mascarenhas por ter aceitado o desafio para ser coorientadora deste trabalho;

A educador Joberto Sérgio Barbosa Martins por ter aceitado o desafio para ser orientadora deste trabalho;

Aos Professores do curso, em especial Galdir Reges e Sérgio Fernandes;

Ao Professor José Vicente Cardoso Santos, pelas horas, ensinos e contribuições impagáveis;

A Professora Marla Dore, pelo acolhimento e incentivo inicial em continuar com a pesquisa;

Ao centro universitário Jorge Amado pelo acolhimento através de seus colaboradores;

A todos que de uma forma ou de outra, contribuíram para a realização desde estudo, o meu muito obrigado.

RESUMO

A evasão universitária certamente aflige diversas instituições acadêmicas particulares de nível superior, e a constante busca de fenômenos que a causa, tem sido foco de muitas pesquisas em educação e tecnologias educacionais no mundo. Assim, a previsão sobre a evasão em unidades de ensino, em especial torna-se importante com foco na formulação de ações que diminuam o impacto desta para as instituições assim como ao estudante em seu desenvolvimento educacional. As técnicas de aprendizado de máquina (*machine learning* - ML) podem auxiliar instituições de ensino e demais áreas educacionais a preverem a evasão. Esse trabalho tem como objetivo geral apresentar uma proposta de método de referência para previsão de evasão utilizando algoritmo de aprendizado de máquina com a técnica de árvore de decisão (*decision tree*). Adota-se uma metodologia de pesquisa híbrida com revisão da literatura de caráter exploratório, explicativo e descritivo do cenário da evasão com itinerário metodológico específico e universo de pesquisa e dados coletados sendo o meio universitário específico. Objetiva-se como resultado a criação de um modelo para a análise de evasão universitária com uso da técnica de árvore de decisão.

Palavras-chave: Educação, evasão, aprendizado de máquina, árvore de decisão.

ABSTRACT

University dropout certainly afflicts several higher-level private academic institutions, and the constant search for phenomena that cause it has been the focus of much research in education and educational technologies globally. Thus, predicting evasion in teaching units, in particular, becomes essential, focusing on formulating actions that reduce its impact on institutions and students in their educational development. Machine learning (ML) techniques can help teaching institutions and other educational areas predict dropout. The general objective of this work is to present a proposal for a reference model for predicting dropout using a machine learning algorithm with the decision tree technique. A hybrid research methodology is adopted with a literature review of an exploratory, explanatory, and descriptive nature of the dropout scenario with a specific methodological itinerary and research universe and collected data being the specific university environment. The objective is to create a model for analyzing university dropouts using the decision tree technique.

Keywords: Education, evasion, machine learning, decision tree.

LISTA DE SIGLAS E ACRÔNIMOS

- ABNT - Associação Brasileira de Normas Técnicas;
- AdaBoost - Algoritmo que combina várias instâncias de um classificador fraco para formar um classificador forte;
- Agglomerative Clustering - Agrupamento Aglomerativo;
- ANDIFES - Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior;
- BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies;
- CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior;
- CEFETs - Centros Federais de Educação Tecnológica;
- CF - Clustering Feature;
- Clusterização - Técnica que agrupa dados não rotulados com base em suas similaridades ou diferenças;
- CNN - Redes Neurais Convolucionais;
- DBSCAN - Density-Based Spatial Clustering of Applications with Noise.
- *Decision Tree* - Algoritmo de classificação e regressão para uso na modelagem preditiva de atributos discretos e contínuos;
- Deep Learning - Aprendizagem profunda;
- Dendrograma - Tipo específico de diagrama ou representação icónica que organiza determinados fatores e variáveis;
- EAD - Ensino a Distância;
- EM - expectation maximization.
- Explotação - Tirar proveito financeiro de uma terra ou área, buscando seus recursos naturais;
- Forms - Formulário Eletrônico da Google;
- Gaussian Naive Bayes - Variação do Naive Bayes que assume que os dados seguem uma distribuição normal (gaussiana);
- GMM - Gaussian Mixture Models;
- Gradient Boosting - Técnica que combina vários modelos de classificação mais fracos em um modelo forte com geração de algoritmo com previsões iniciais com ajustes subsequentemente;
- Hierarchical Clustering - Agrupamento Hierárquico;
- IA - inteligência artificial;
- IES - Instituições de Ensino Superior;
- IETS - Instituto de Estudos do Trabalho e Sociedade;
- IFs - Institutos Federais de Educação, Ciência e Tecnologia;
- INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira;
- k clusters - k é um número pré-definido, minimizando a soma dos quadrados das distâncias entre os pontos e seus respectivos centroides;
- Kernel SVM - Máquinas de Vetores de Suporte com Kernel;
- k-means - Algoritmo que divide os dados em k clusters com minimização da soma dos quadrados das distâncias entre os pontos e seus respectivos centroides;
- LDB - Lei de Diretrizes e Bases da Educação Nacional;
- Logistic Regression - Regressão Logística;
- Machine Learning - Disciplina da área da Inteligência Artificial que, por meio de algoritmos, dá aos computadores a capacidade de identificar padrões em dados massivos e com isso faz-se previsões (análise preditiva);
- Matplotlib - Submódulo do Matplotlib com foco na criação de gráficos;
- Mean Shift - Técnica de análise matemática de espaço no deslocamento médio;
- Naive Bayes - São classificadores probabilísticos baseados na aplicação do teorema de Bayes;
- OCDE - Organização para a Cooperação e o Desenvolvimento Econômico. Relatórios Econômicos;
- OPTICS - Ordering Points To Identify the Clustering Structure;

- Outliers - Valor aberrante ou valor atípico;
- Python - Linguagem de programação de alto nível;
- Random Forest - É um método de florestas aleatórias ou florestas de decisão aleatórias para aprendizado conjunto, classificação, regressão e outras tarefas;
- Randomized *Decision Tree*s - Árvores de Decisão Estocásticas;
- Regressão Bayesiana - Regressão que reflete a estrutura bayesiana, ou seja, a elaboração de hipóteses, testá-las com experiências e observações e reajustar crenças iniciais de acordo com as evidências (processo de inferência descrito pelo teorema de Bayes);
- Regressão de Árvore de Decisão - Método de aprendizado supervisionado;
- Regressão de Floresta Aleatória - Árvore de decisão que faz perguntas e acaba prevendo uma classe ou atribuindo uma classe (ou seja, atribuindo uma resposta à pergunta);
- Regressão Linear Múltipla - Modelo de análise usada quando se modela a relação linear;
- Regressão Linear Simples - Técnica de análise de dados que prevê o valor de dados desconhecidos usando outro valor de dados relacionado e conhecido para uma reta;
- Regressão Logística Multinomial - Trata de um modelo de regressão logística que permite que a variável categórica dependente apresente mais de duas categorias;
- Regressão Polinomial - Técnica de análise de dados que prevê o valor de dados desconhecidos usando outro valor de dados relacionado e conhecido para polinômios de ordens variadas;
- Regressão por Redes Neurais - Método de aprendizado supervisionado que requer um conjunto de dados marcado com previsão de valor numérico;
- Regressão Ridge e Lasso - Overfitting - Regressão com capacidade de lidar com multicolinearidade com escolha de coeficientes em direção a zero;
- Robustez a outliers - Formas robustas de regressão que minimizam os erros medianos de mínimos quadrados em vez da média (também chamada de regressão robusta);
- Scikit-learn - Biblioteca da linguagem Python desenvolvida especificamente para aplicação prática de machine learning;
- SciPy - Pacote principal de rotinas científicas em Python;
- SESu/MEC - Secretaria de Educação Superior do Ministério da Educação e Desporto;
- SOM - Self-Organizing Map;
- Spectral Clustering - Agrupamento Espectral;
- SVM - Regressão de Máquina de Vetores de Suporte ou Support Vector Machines;

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Desde a aprovação da Lei nº 9.394/96, que descreve as Diretrizes e Bases da Educação Nacional, o Estado brasileiro estabeleceu como um dos princípios e fins da educação a garantia do padrão de qualidade do ensino público ofertado (Brasil, 1988).

Nesse contexto, a introdução das políticas avaliativas educacionais, em larga escala, justificou-se em razão do processo de expansão da escolarização no Brasil bem como o acompanhamento dos processos de desistência ou abandono de alunos em todos os níveis educacionais, ou seja, da evasão (Banco Mundial, 2017).

Nesse caso, é preciso compreender a educação escolar como fenômeno social, articulado às políticas econômicas, expressão das diferentes forças sociais que procuram se legitimar de maneira hegemônica no espaço educacional; vale dizermos: a escola é um espaço de disputas de teorias, de métodos e de práticas e portanto a evasão desse ambiente é altamente prejudicial ao pais (Almodóvar-González; Marugán-de-Miguelsanz, 2021), não apenas nos seus aspectos acadêmicos mas também nos seus aspectos financeiros, sociais não somente no Brasil mas em todo o mundo (OCDE, 2018) considerando-se então a evasão como um dos grandes problemas da baixa efetividade dos processos educativos no Brasil.

De acordo com Caro (2013), a evasão escolar, ou em meios escolares de todos os níveis, é definida como sendo "... o abandono voluntário da escola antes da conclusão do ensino obrigatório ou da formação desejada", e a isso atribui-se graves consequências para o indivíduo, incluindo redução da qualidade de vida e oportunidades de emprego limitadas (Almodóvar-González; Marugán-de-Miguelsanz, 2021).

As diretrizes educacionais delineadas anteriormente enfrentam um desafio preocupante: a evasão escolar, caracterizada pelo abandono prematuro dos alunos em suas jornadas educacionais. Essa problemática representa uma ameaça significativa ao progresso almejado no campo da educação. É uma preocupação crescente em todo o mundo e em todos os níveis da educação, da escola, nos seus níveis fundamentais as faculdades e universidades em todos os seus cursos. Trata-se de um dos grandes problemas que afetam negativamente a educação (Neri, 2009).

Para se obter um bom resultado, nos aspectos acadêmicos e sociais, um dos possíveis caminhos é o investimento na manutenção dos processos educacionais com a redução das respectivas evasões.

Para Gaioso (2005) a evasão escolar é compreendida como um intrincado fenômeno de natureza social, caracterizado pela interrupção do percurso educacional de um aluno em seu ciclo de estudos, e se pode dizer que as suas causas podem ser variadas, o que inclui a falta de motivação, problemas familiares, dificuldades financeiras, entre outros ao tempo em que evidencia que a evasão escolar tem consequências significativas tanto para os indivíduos quanto para a sociedade como um todo, pois pode limitar as oportunidades futuras e aumentar a desigualdade social (Almodóvar-González; Marugán-de-Miguelsanz, 2021).

De acordo com Silva, Cabral e Pacheco (2020), a evasão escolar traz uma série de consequências negativas tanto para o estudante quanto para a instituição. O estudante corre o risco de não adquirir uma formação adequada e, ao mesmo tempo, perde tempo e recursos financeiros.

No entanto, a evasão também pode ser vista como uma escolha consciente do estudante em busca de outras formações ou objetivos pessoais e, por outro lado, a instituição sofre com a perda de eficiência decorrente da evasão (Almeida; Ramalho, 2020).

A tecnologia tem apoiado diversas áreas e pode ser utilizada no contexto da previsão da evasão escolar. A Inteligência Artificial (IA), por exemplo, desempenha um papel crucial nesse sentido. A aplicação de modelos de aprendizado de máquina (ML) (Amorim; Barone; Mansur, 2008) para prever a evasão pode ser uma solução eficaz, permitindo que as escolas implementem medidas preventivas e personalizadas para cada aluno.

Essa abordagem tem sido amplamente adotada em outros domínios, evidenciando sua utilidade e potencial. A utilização da ML no combate à evasão escolar pode contribuir para a melhoria da qualidade da educação e para a redução desse problema socialmente relevante (Almodóvar-González; Marugán-de-Miguelsanz, 2021).

1.2 OBJETO DA PESQUISA

Com base neste contexto, faz-se necessário que as instituições possam prever a evasão com base em indicação do perfil estudantil para que possa minimizar os impactos sociais e financeiros investidos no processo de formação educacional.

Para isto a tecnologia está permitindo que sejam desenvolvidos novos métodos para prevenir a evasão escolar. O uso do aprendizado de máquina, pode ser uma ferramenta valiosa na identificação de estudantes em risco de evasão escolar (Gómez *et al*. 2018).

Diversos algoritmos e técnicas para predição de evasão escolar estão sendo desenvolvidas com objetivo de identificar os alunos em risco de evasão (Costa *et al.*, 2017), que compararam as eficácias de técnicas de Mineração de Dados Educacionais para identificar alunos em risco (Adejo; Connolly, 2017). Porém, as pesquisas realizadas com a utilização de Aprendizado de máquina se beneficiam de modelos estatísticos para previsão de risco de algum evento ocorrer (Silva; Bezerra; Brasil, 2018).

O Aprendizado de máquina permite que sejam analisados grandes conjuntos de dados para identificar padrões e tendências que possam ser usados para prever e compreender a evasão (Harrison, 2020). Uma das principais abordagens utilizadas é a análise de mineração de dados, que busca extrair informações pertinentes em uma grande quantidade de dados (Domingos, 2017).

Além disso, o aprendizado profundo (*Deep learning*) tem sido utilizado com sucesso para prever a evasão ao tempo em que as redes neurais podem ser utilizadas para prever a evasão com base em dados demográficos, de desempenho acadêmico e de participação dos alunos (Zhang *et al.*, 2019), dessa forma a tecnologia está sendo usada para apoiar a solução desse problema.

Desta forma, esta pesquisa aborda a problemática da evasão em cursos superiores considerando a diversidade de tecnologias disponíveis. O objetivo central é compreender o conceito de evasão e explorar as diversas perspectivas oferecidas pelas técnicas de mineração de dados educacionais.

Isso nos permitirá identificar alunos em risco de evasão por meio da aplicação de algoritmos de Aprendizado de Máquina. É importante destacar que a escolha do método dependerá das características específicas do problema e das limitações de cada abordagem tecnológica."

1.3 OBJETIVOS: GERAL E ESPECÍFICOS

Nesta seção, serão apresentados os objetivos, geral e específicos, que foram utilizados para nortear esta pesquisa.

Este trabalho tem como objetivo geral apresentar um método de referência para previsão de evasão através da análise de algoritmos de *Aprendizado de máquina*, como objetivos específicos tem-se:

a) Realizar levantamento bibliográfico, em atual estado da arte, sobre os fatores que influenciam a evasão escolar e algoritmo de decisão em árvore;
b) Identificar a percepção de gestores da educação em IES sobre a evasão de alunos;
c) Propor método de referência para previsão evasão em IES;
d) Validar o método proposto;

1.4 JUSTIFICATIVA

A evasão é um problema preocupante em todo o mundo, especialmente em países em desenvolvimento, onde as taxas de abandono escolar são elevadas. Esse fenômeno pode ter consequências negativas para o indivíduo, a sociedade e a economia, acarretando menor produtividade, maior desigualdade social e redução do crescimento econômico.

Nesse contexto, a criação de um método de Aprendizado de máquina para previsão de evasão escolar apresenta-se como uma solução valiosa, pois permite a identificação precoce dos alunos em risco de abandonar a escola antes que isso aconteça (Harrison, 2020).

Ao antecipar esse cenário, as instituições de ensino podem implementar intervenções apropriadas, auxiliando os alunos a permanecerem na escola e a alcançarem o sucesso acadêmico (Domingos, 2017).

Contudo, o desenvolvimento de tais métodos não é tarefa simples, requerendo a análise de uma vasta quantidade de dados, incluindo informações demográficas, acadêmicas e comportamentais dos alunos.

Além disso, é fundamental garantir que os métodos sejam justos e imparciais,

de modo a evitar a discriminação contra determinados grupos de estudantes (Adekitan; Salau, 2019).

Neste sentido, a contribuição significativa deste trabalho está na proposição de um método de referência para previsão de evasão escolar. O método de referência agrega os parâmetros mais relevantes já utilizados em outros métodos e define o algoritmo a ser empregado, facilitando assim a tarefa de instituições que desejem criar seus próprios métodos de predição.

A importância do método de referência reside na sua capacidade de proporcionar às instituições de ensino uma base sólida para a criação de métodos específicos de evasão. A complexidade desse processo reside no desafio de selecionar as variáveis (parâmetros) mais adequadas e escolher os algoritmos e ferramentas mais apropriados. Nosso trabalho se destaca por realizar uma extensa pesquisa de método existentes, estudar diferentes algoritmos e conduzir pesquisas junto a coordenadores de instituições educacionais.

Assim, o método de referência que propomos é uma compilação desses esforços, permitindo que outras instituições não precisem reproduzir os mesmos passos.

Além de fornecer um método de referência, esta dissertação também se apresenta como uma contribuição significativa para a área de gestão educacional, visando a excelência na gestão e na solução de problemas enfrentados cotidianamente.

A escolha do tema é uma forma de cooperar com a instituição onde atuo, bem como com os cursos de tecnologia e o setor educacional em geral, com o propósito de minimizar o fenômeno da evasão que afeta as instituições de ensino superior (Durhan; Sampaio, 2001) de forma recorrente (Adekitan; Salau, 2019).

Em síntese, este trabalho visa proporcionar uma ferramenta valiosa às instituições de ensino, oferecendo um método de referência para a previsão de evasão escolar. Esse método é fruto de um extenso estudo e pesquisa, tornando-se uma base sólida para a criação de método específicos de evasão em diferentes contextos institucionais.

Acreditamos que essa abordagem possa auxiliar as instituições no combate à evasão, contribuindo para a sua sustentabilidade econômico-financeira e o aprimoramento da gestão educacional como um todo.

1.5 METODOLOGIA

Abaixo segue descritivo da metodologia aplicada nesta pesquisa perpassando pela técnica da pesquisa adotada, itinerário metodológico e auto-consistência e completude da pesquisa.

1.5.1 O problema da pesquisa e questão focal

O problema da pesquisa perpassa por levantamento da questão da necessidade de um método de referência em *Aprendizado de máquina* para a análise da evasão em IES.Com isso tem-se a seguinte questão de pesquisa: é possível especificar um método de detecção precoce de evasão escolar que possa ser utilizado como referência para universidades que desejem utilizar ML para apoiar o combate a evasão?

1.5.2 Técnicas adotadas na pesquisa

Nesta seção, apresentaremos a metodologia adotada na pesquisa, buscando simplificar a exposição conforme sugerido. A pesquisa realizada é uma combinação de abordagens aplicada e exploratória.

A pesquisa aplicada utilizou conhecimento pré-existente para propor um método de referência, enquanto a pesquisa exploratória se baseou em métodos de referência já existentes em outras áreas, para compreender e reunir conhecimento, culminando no desenvolvimento do nosso próprio método de referência.

A metodologia adotada pode ser classificada como uma pesquisa, exploratória e descritiva (Gil, 2021). Realizamos uma pesquisa exploratória para obter percepções dos coordenadores sobre o processo de evasão, utilizando um questionário eletrônico (Forms).

Nesse cenário, essa pesquisa é exploratória (Gil, 2021), ou seja, vamos prospectar opiniões dos coordenadores, sobre o processo de evasão com instrumento de prospecção, o questionário eletrônico Forms, e, com isso, treinar a rede com a técnica Árvore de Decisão.

1.5.3 Itinerário metodológico

Para responder ao questionamento principal da pesquisa, adotaremos um itinerário mais simplificado, seguindo a abordagem utilizada na seção anterior. Nosso objetivo é apresentar o passo a passo utilizado para o desenvolvimento do trabalho, de forma clara e compreensível.

Inicialmente, forneceremos uma visão geral do processo adotado para lidar com o conjunto de variáveis associadas ao processo decisório de evasão no ensino superior. Para isso, utilizaremos a técnica da Árvore de Decisão como foco, buscando identificar e analisar os fatores que influenciam a evasão dos alunos.

Em seguida, descreveremos o método de previsão de evasões baseado nos resultados propostos por Durhan e Sampaio (2001), adaptando-o ao contexto específico da nossa pesquisa.

Esse método será fundamental para o treinamento e teste do processo decisorial preditivo, que utilizará um questionário de pesquisa para a coleta de dados (Adejo; Connolly, 2017). O objetivo dessa etapa é criar uma base de dados consistente para o treinamento do método de previsão de evasão.

Para tornar a metodologia mais compreensível, também incluiremos figuras e diagramas ilustrativos ao longo da explicação, permitindo que os leitores visualizem melhor o processo adotado. Essas ilustrações servirão como ferramentas para facilitar a compreensão e acompanhar a evolução das etapas.

1.5.4 Classificação

Não obstante, será também uma pesquisa bibliográfica, em revisão de literatura dos temas afins, ou seja, da evasão no meio universitário brasileiro e baiano bem como novas versões de explicações da técnica em uso, e tudo isso, com caráter experimental pois vamos prospectar dados no universo da pesquisa que será o meio universitário do Centro Universitário específico.

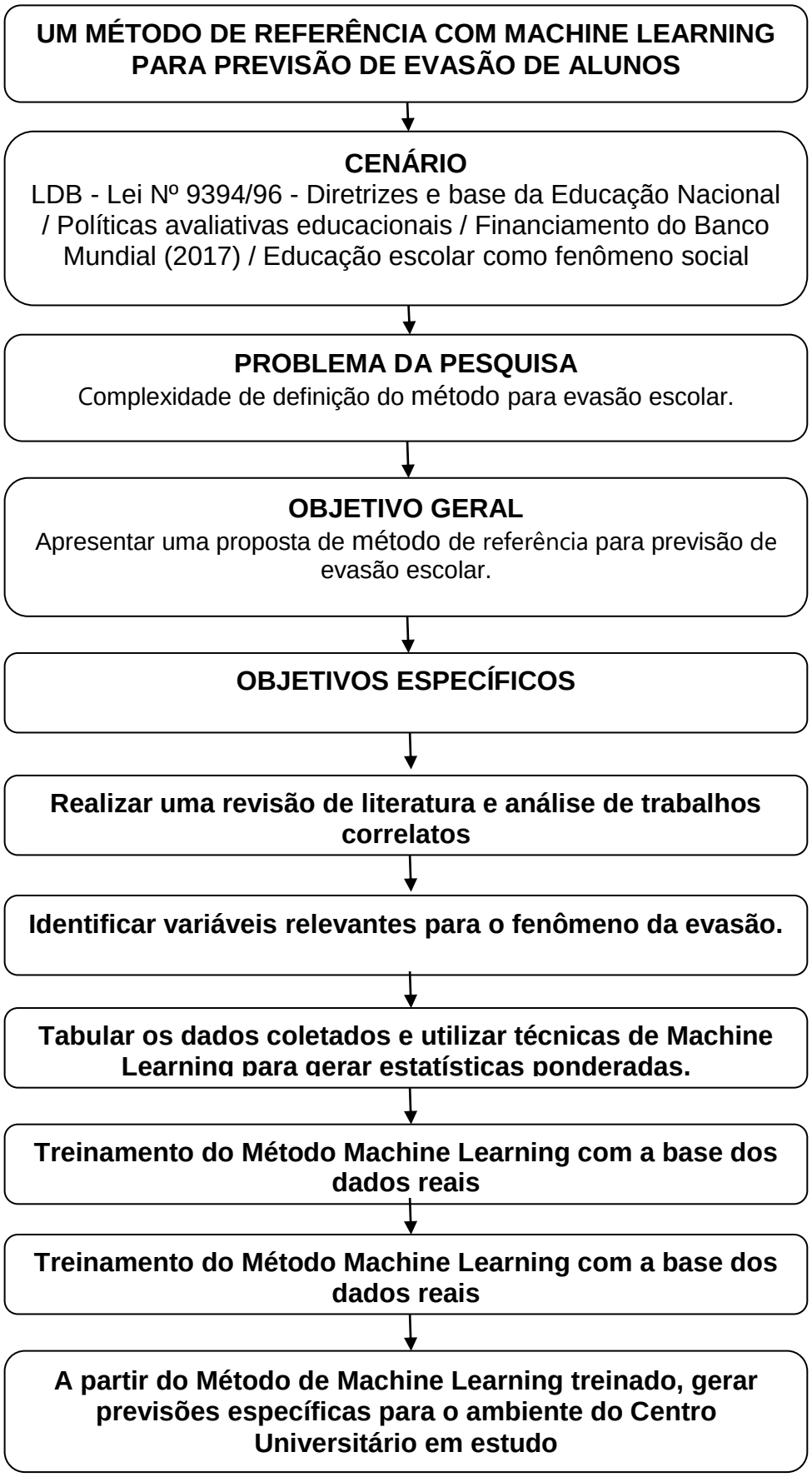
Dessa forma trata-se então de uma metodologia híbrida com revisão de literatura de caráter exploratório, explicativo e descritivo do cenário da evasão do Centro Universitário específico com itinerário metodológico específico que será composto das seguintes etapas no seu processo de elaboração:

1) Realizar uma revisão de literatura com foco na evasão escolar em ambientes universitários e em técnicas algorítmicas de mineração de banco de dados. E análise de trabalhos correlatos que abordem outros métodos de evasão, a fim de situar o trabalho em relação às abordagens anteriores;
2) Identificar variáveis relevantes para o fenômeno da evasão com lastro na literatura pertinente;
3) Tabular os dados coletados e utilizar técnicas de Aprendizado de máquina para gerar estatísticas ponderadas, a fim de treinar o método. Utilizar a técnica Árvore de Decisão para identificar eventuais correlações entre as variáveis e o fenômeno da evasão;
4) Treinamento do método *Aprendizado de máquina* com a base dos dados reais e com o uso da técnica Árvore de Decisão;
5) A partir do método de Aprendizado de máquina treinado e validação do trabalho.

1.5.5 Auto-consistência e completude da pesquisa

Com o exposto pode-se afirmar, agora, que o cenário motivador da pesquisa, o problema da pesquisa, os seus objetivos, geral e específicos, bem como as justificativas estão em completa consonância, conforme pode-se verificar na figura integrativa a seguir:

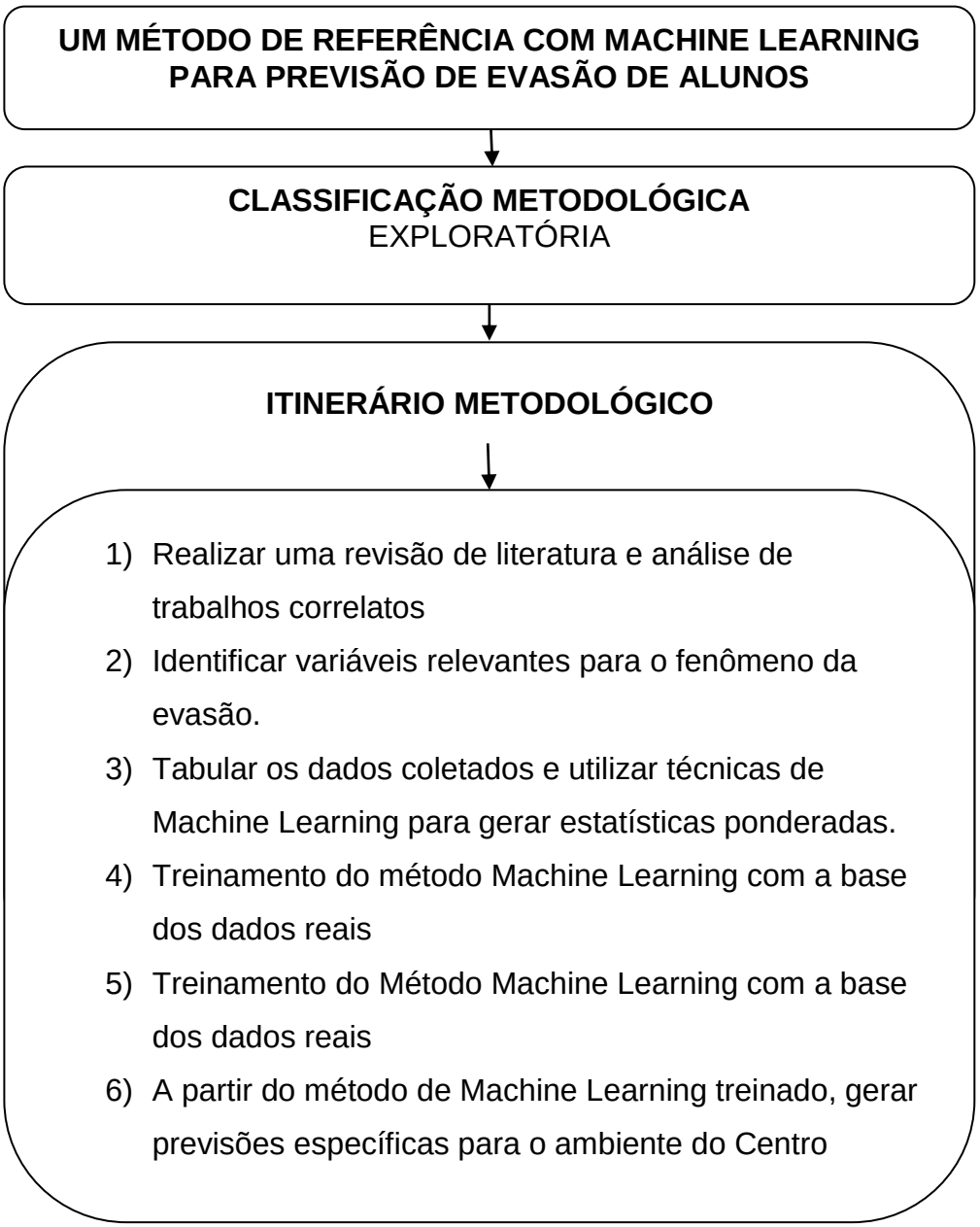
Figura 1 - Cenários e objetivos



Fonte: do autor (2023).

Além disso a metodologia e o itinerário metodológico também podem ser associados da seguinte forma:

Figura 2 - Metodologia e Itinerário Metodológico



Fonte: do autor (2023).

1.6 ESTRUTURA DA DISSERTAÇÃO

Para estruturar a dissertação, foi realizada as seguintes divisões em capítulos organizados da seguinte forma: no capítulo 1 tem-se a apresentação de

todo trabalho em conformidade com a norma ABNT 15.287, como sendo projeto de pesquisa, relatando seu texto de aproximação, cenários, objeto de pesquisa, objetivos, geral e específicos, justificativa e metodologia.

No capítulo 2, realizamos uma revisão da literatura relacionada às temáticas em questão. Começamos abordando os conceitos de evasão em ambientes acadêmicos e universitários, bem como os princípios fundamentais da inteligência computacional e suas ramificações. Além disso, exploramos as técnicas associadas a esses conceitos, destacando as abordagens mais recentes e avançadas no estado da arte.

Dentro desse contexto, dedicamos atenção ao aprendizado de máquina, seus conceitos-chave e aplicações relevantes. Ademais, apresentamos os principais algoritmos utilizados na análise de dados, com um foco particular na técnica da Árvore de Decisão. Detalhamos suas características distintivas e nuances essenciais, destacando sua importância no contexto da nossa temática de estudo.

No capítulo 3, é apresentada uma proposta de método de referência que utiliza Aprendizado de Máquina para prever a evasão de alunos. Esta proposta está baseada na literatura atualizada sobre o tema, bem como nas técnicas de coleta de dados e tabulação, que incluem o uso da Árvore de Decisão.

O Capítulo 4, denominado "Avaliação do Método Proposto", desempenha um papel crucial ao concluir nossa pesquisa, apresentando também orientações para possíveis futuros trabalhos.

Adicionalmente, este capítulo inclui os Apêndices A e B, que enriquecem os dados e informações ao longo de todo o estudo. Essa estrutura composta por cinco capítulos foi cuidadosamente elaborada para assegurar uma divisão clara e lógica do conteúdo da dissertação.

O Capítulo 5, compreende os comentários finais sobre a pesquisa realizada e oferece sugestões para trabalhos futuros. Além disso, são anexados apêndices A e B que complementam as informações fornecidas ao longo da pesquisa.

Essa estrutura organizada em cinco capítulos proporciona uma clara e coerente divisão do conteúdo da dissertação.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta fundamentação, será abordado os conceitos e teorias fundamentais que sustentam a pesquisa, discutindo a importância dos fundamentos teóricos para embasar a compreensão do problema da evasão acadêmica.

Logo após, serão abordos os cenários e elementos motivacionais que influenciam a permanência dos estudantes nas instituições de ensino, assim como os fatores que motivam os alunos a continuar seus estudos e os desafios que podem levar à evasão.

Em seguida, será aprofundado o conhecimento sobre os ambientes acadêmicos e sua relação com a evasão. Investigaremos as causas desse fenômeno, analisando as características e nuances presentes no ambiente universitário.

Entraremos no campo da Inteligência Computacional, explorando os princípios e capacidades dessa área. Será foca em Aprendizado de máquina, discutindo a hierarquia do aprendizado e as características dos diferentes tipos de aprendizado, como supervisionado, não supervisionado, semi-supervisionado e por reforço (Amorim; Barone; Mansur, 2008).

Por fim, destacaremos a relevância do desenvolvimento de um método como uma abordagem adequada para lidar com a evasão e sua aplicação no método de referência proposto.

Ao fim deste capítulo, teremos uma base sólida de conhecimento teórico que embasará a análise e o desenvolvimento de soluções para a problemática da evasão escolar e acadêmica, utilizando a inteligência computacional.

2.1 CENÁRIOS E ELEMENTOS MOTIVACIONAIS

O ensino superior, nos seus moldes modernos, tem seu surgimento no Brasil em 1960 e ganha realmente notoriedade a partir de 1980 com a fundação de universidades privadas com a finalidade de promover o crescimento econômico do país inclusive com novos avanços na década de 90 com a implementação da Lei de Diretrizes e Bases da Educação - LDB, a Lei 9394 de 20 de dezembro de 1996.

Conforme (INEP, 2021), O Censo de 2020, informa que possuímos 2.456 instituições de educação superior, onde 87,6% das instituições são privadas, e

detém em suas instalações um total de 8,6 milhões de alunos matriculados (Durhan; Sampaio, 2001).

As instituições de ensino precisam estar prontas para poderem se relacionar com uma geração mais atualizada e com os meios de comunicação atuais, especialmente a Internet, o acesso à informação é instantâneo e os alunos têm facilidade em buscar conhecimento por meio das tecnologias disponíveis (Faria, 2004).

De acordo com a visão apresentada por Moran (2007), os métodos educacionais devem enfatizar a criação conjunta de conhecimentos, utilizando a tecnologia como mediadora. Nesse contexto, o professor assume um papel ativo e participativo, intermediando e orientando o processo de construção do conhecimento.

Desta forma, entre muitos acontecimentos e marcos históricos a educação foi se transformando e expandindo para níveis mais avançados, chegando ao ambiente universitário, um espaço de aprendizagem e pesquisa onde estudantes e professores reunidos buscam pelo conhecimento e inovação (Amorim; Barone; Mansur, 2008).

2.2 AMBIENTES ACADÊMICOS E A EVASÃO

Será abordado o tema da evasão nos ambientes acadêmicos com o foco de entendermos sobre este fenômeno. Primeiramente, é discutido as causas que levam à evasão escolar e acadêmica, destacando os principais fatores que contribuem para essa problemática. Em seguida, será focado no ambiente universitário, explorando suas características e nuances específicas relacionadas à evasão.

Por fim, será examinado as tipologias da evasão, identificando os diferentes padrões e perfis dos estudantes que abandonam os estudos. Ao compreender esses aspectos, busca-se obter uma visão abrangente desse fenômeno e propor possíveis soluções para reduzir a evasão nos ambientes acadêmicos.

2.2.1 Causas da evasão escolar e acadêmica

Os estudos revisados na literatura apontam que as causas da evasão escolar estão associadas a comportamentos variados por parte dos estudantes, os quais

são influenciados pelas experiências de vida individuais, pelas situações vivenciadas compartilhadas no ambiente acadêmico em que se encontram.

Conforme Grayson e Grayson (2003), foi realizado um estudo que tem como base a retenção e evasão, no Canadá que mostra as principais diferenças nas taxas de evasão a depender do ambiente acadêmico avaliado, ou seja, a evasão é menor em campos profissionais, como Educação, Engenharia, Gestão de Empresas, Direito, entre outros, do que em áreas como das artes e da ciência.

Analisando um cenário local entre as instituições brasileiras, podemos encontrar dados que afirmar que o ambiente tem passado por transformações que impactam diretamente na decisão do aluno em optar por evadir.

2.2.2 Ambiente universitário: características e nuances da evasão

O investimento realizado pelas instituições públicas e privadas com a finalidade de promover a educação no país ultrapassa a casa dos bilhões, entretanto, o retorno sobre recurso não traz resultado aos gestores.

Por outro lado, houve um crescimento na abertura de novos postos de educação conforme dados publicados pelo instituto nacional de estudos e pesquisas educacionais Anísio Teixeira INEP no último censo, há 313 IES públicas e 2.261 IES privadas no Brasil.

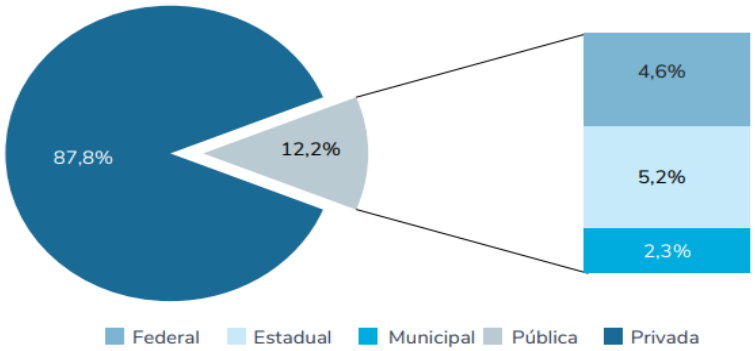
No que diz respeito a Instituição de Nível Superior (IES) públicas: 42,8% são do estado (134 IES); 38,0% são mantidas pelo governo federal (119); e 19,2% são de origem municipal (60); mais da metade das universidades são instituições públicas, representando 55,4% do total. Entre as instituições de ensino superior (IES) privadas, as faculdades são a maioria, correspondendo a 81,0% (Durhan; Sampaio, 2001).

Cerca de três quintos (3/5) das IES federais são universidades, enquanto 34,5% são compostas pelos Institutos Federais de Educação, Ciência e Tecnologia (IFs) e pelos Centros Federais de Educação Tecnológica (Cefets), de acordo com dados do INEP (2021).

Figura 3 - Instituições de educação superior, por organização acadêmica e categoria administrativa 2021

Ano	Total	Universidade		Centro Universitário		Faculdade		IF e Cefet	
		Pública	Privada	Pública	Privada	Pública	Privada	Pública	Privada
2021	2.574	113	91	12	338	147	1.832	41	N.A.

Fonte: Elaborada pela Deed/Inep com base em dados do Censo da Educação Superior.
Nota: N.A.(Não se aplica).



Fonte: Censo da Educação Superior: notas estatísticas INEP (2021).

É necessário entender o perfil do aluno que chega às instituições de ensino superior e perceber que seus perfis são cada vez mais heterogêneo e abrange diversas camadas sociais da população. Este cenário mostra que um quantitativo significativo de alunos estão cada vez mais sendo impactados em seu processo decisório pendendo a evadir das instituições (Adekitan; Salau, 2019).

A evasão universitária é um problema que pode comprometer não apenas o desenvolvimento dos estudantes, mas também a economia do país. Isso porque a formação de ensino superior é vista como uma maneira de inserir indivíduos no mercado de trabalho (Barreiro; Terribili, 2007).

No entanto, outros problemas relacionados a falta de motivação, problemas financeiros, problemas de saúde mental, dificuldades acadêmicas podem levar à evasão, o que evidencia a necessidade de criar um ambiente de aprendizagem inclusivo, que impeça tais fatos.

2.2.3 O surgimento das primeiras instituições de ensino superior no Brasil

Os Jesuítas foram responsáveis pela fundação do primeiro estabelecimento de Ensino Superior do Brasil sediado na Bahia, então sede do Governo Geral, em 1550, segundo Cunha (2001).

Oficialmente, a criação da primeira universidade aconteceu em 1909, em Manaus, no Amazonas, como resultado da iniciativa de grupos privados, durante o

breve período de florescimento da indústria borracha, que terminou em 1926.

A Universidade do Rio de Janeiro (hoje conhecida como Universidade Federal do Rio de Janeiro) foi estabelecida em 1920, tornando-se a primeira instituição de ensino superior do Brasil a obter permanentemente o status de universidade, fruto da união das Escolas Politécnica, Medicina e Direito, do Governo Federal.

No ano de 1927, foi fundada a instituição de nível superior a Universidade de Minas Gerais, que ocorreu através da fusão de outras faculdades como Faculdades de Engenharia, Faculdade de Direito, Faculdade de Odontologia, Faculdade de Farmácia e Faculdade de Medicina que eram existentes Belo Horizonte.

Uma das características do ensino superior na década de 1960 nas questões que envolvia o debate das expansões do ensino era que ele era frequentado pela elitizado, mesmo que fosse ofertado pelas instituições públicas, Baggi (2011).

Entretanto na década de 1970 o processo de expansão e popularização do ensino superior foi iniciado com um havendo importante para a época que foi chamado de reforma universitária. Tal reforma tinha como tarefa analisar e conferir maior eficiência e modernização as instituições universitárias e fomentar o ingresso de jovens de classe média e trabalhadores assalariados no ensino superior.

A Lei nº 5.540/68 (Brasil, 1968), que estabeleceu a reforma universitária, foi ampla e profunda, rompendo com a composição vigente até então de estrutura dos cursos superiores no Brasil (Schwartzman, 2001).

Antes da reforma de 1968, a Educação Superior era composta, basicamente, por escolas profissionais independentes, com pouca ou nenhuma ênfase em pesquisa.

Entre as medidas adotadas com a citada lei, cabe destacar a criação do sistema de institutos básicos e a instituição do departamento como unidade mínima de ensino e pesquisa, a alteração do vestibular, e disposição decretando os sistemas de créditos e de semestralidade.

De acordo com Adachi (2010), o fato do aumento do ingresso no Ensino Superior deve-se a 2 (dois) fatores principais, a saber:

1) aumento no nível de escolarização da população, com consequente aumento do número de concluintes do Ensino Médio; e

2) facilidades proporcionadas pelas políticas de flexibilização e regulamentação da Educação Superior, por parte dos governos.

Diante de tal aumento, o perfil do aluno que chega às instituições de ensino superior torne-se cada vez mais heterogêneo e abrange diversas camadas sociais da população.

Como defende Mello (2007), os novos alunos que são atendidos contemplam ainda mais as mais diversas classes da população que não exclui a relativa situação de perpetuação das desigualdades e as oportunidades de escolarização e no longo prazo, se verifica a quantidade e o perfil dos alunos que são concluintes do Ensino Superior.

2.2.4 Evasão e suas tipologias

A evasão é considerada uma das maiores preocupações do ensino superior, principalmente, Segundo Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior - ANDIFES (1996), é crucial que as instituições públicas assegurem resultados relevantes e validem a formação dos estudantes para a inserção no mercado de trabalho.

Flores (2017) esclarece que a palavra evasão vem do latim *evasio* e significa fuga, saída, abandono, fracasso, insucesso, mas a terminologia adotada varia conforme o autor, porém, todos que iniciam um curso têm o mesmo propósito: o de finalizar um curso.

De acordo com Silva, Cabral e Pacheco (2020), Costa e Gouveia (2018) e Prestes e Fialho (2018), não existe concordância da literatura sobre o conceito de evasão.

A partir dessa avaliação, o estudo de Costa e Gouveia (2018) forneceu um quadro com definições para retenção e persistência, pois eles mostraram que, para autores que diferenciam esses conceitos, há diversas definições, já para os autores que não diferenciam ou os autores que apenas definem retenção não há consenso sobre os termos.

Outra questão que os autores mostram é que a maior parte das definições é proveniente de pesquisas internacionais, sendo principalmente norte-americanas

(Costa; Gouveia, 2018).

A Andifes (1996) definiu a evasão no ensino superior como sendo a saída definitiva do aluno do curso de origem sem concluí-lo. Essa evasão pode ser dividida em três tipos, conforme destacam Andifes (1996) e Prestes e Fialho (2018):

a) Evasão do curso: o estudante desliga-se do curso por abandono, deixando de fazer matrícula, desistir de maneira oficial, transferir-se ou alterar o curso, ou mesmo ser excluído de acordo com as normas institucionais;

b) Evasão da instituição: o estudante desliga-se da instituição em que está matriculado;

c) Evasão do sistema: o estudante desliga-se de forma definitiva ou temporária do ensino superior;

Ao analisar criticamente os dados, torna-se evidente a necessidade de alocar recursos para combater ou, até mesmo, reduzir drasticamente a taxa de evasão de estudantes nas instituições de ensino.

De acordo com Andriola (2004), ao notarmos o aumento constante no número de matrículas no Ensino Superior, fica nítido que há um crescimento significativo no número de estudantes matriculados nas instituições. No entanto, a garantia da permanência desses alunos até a conclusão dos cursos não está assegurada.

De acordo com o INEP (2021), em 2021, foram ofertadas um pouco mais de 22,6 milhões de vagas nos cursos de nível superior, e destas 74,5% vagas novas e 25,2% vagas remanescentes.

Sobre as ofertas, podemos observar que a rede privada ofereceu 96,4% do total de vagas em cursos de nível superior no ano de 2021, por outro lado as instituições da rede pública corresponderam a 3,6% das vagas ofertadas pelas IES e de todas as vagas de caráter remanescente, um total de 97,0% foram ofertadas por IES da rede privada conforme visto na imagem abaixo.

Tabela 1 - Número de vagas de cursos de graduação, por tipo de vaga e categoria administrativa 2021

Categoria Administrativa	Vagas de Cursos de Graduação			
	Total Geral de Vagas	Vagas Novas Oferecidas	Vagas de Programas Especiais	Vagas Remanescentes
Total Geral	**22.677.486**	**16.884.427**	**85.851**	**5.707.208**
Pública	827.045	646.844	6.552	173.649
Federal	491.155	379.125	4.409	107.621
Estadual	229.254	185.282	905	43.067
Municipal	106.636	82.437	1.238	22.961
Privada	21.850.441	16.237.583	79.299	5.533.559

Fonte: Censo da Educação Superior: notas estatísticas INEP (2021).

Das ofertas ocorridas no ano de 2021, apenas 20,2% da oferta conseguiu ser preenchida, e as vagas remanescentes ofertadas pelas instituições apenas 9,0% das foram preenchidas no mesmo período.

O INEP (2021), também informa em sua publicação que 78,2% das novas vagas ofertadas nas instituições da rede federal de ensino foram ocupadas em 2021. Já em relação a rede de ensino estadual teve apenas 26% de preenchimento e na rede feral 86 mil remanescentes não foram preenchidas no ano de 2021.

2.2.5 Indicadores na educação superior no Brasil

Segundo Costa e Gouveia (2018), os estudos sobre evasão no Brasil começaram nos anos de 1995 e 1996 por meio de uma comissão especial pela então Secretaria de Educação Superior do Ministério da Educação e Desporto (SESu/MEC).

Essa comissão avaliou boa parte das instituições federais de ensino superior quanto aos índices de diplomação, de retenção e de evasão dos cursos de graduação (Costa; Gouveia, 2018).

Desde então, o Brasil tem sofrido intensas transformações na educação superior, com a publicação da Lei Geral de Diretrizes e Bases da Educação Nacional (LDB), que incentivou a expansão de matrículas por meio do crescimento das

instituições, de cursos e de vagas, dando autonomia para as instituições públicas de ensino elaborarem, aprovarem e executarem planos de investimentos, de ações e orçamentários, criando, organizando e extinguindo cursos e programas de educação (Brasil, 1996).

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) atua em três frentes no país: Avaliação e Exames Educacionais, Gestão do Conhecimento e Estudos Educacionais e Pesquisas Estatísticas e Indicadores Educacionais.

Nesta última frente, o INEP realiza o Censo da Educação Superior e dispõe de Indicadores de Fluxo da Educação Superior, os quais foram utilizados nesta pesquisa.

O Censo da Educação Superior é um instrumento de pesquisa realizado anualmente sobre as instituições de ensino superior, de cursos de graduação e sequenciais, alunos e docentes.

Os dados utilizados no censo são sobre infraestrutura das instituições, vagas oferecidas, candidatos, matrículas, ingressantes, concluintes e docentes e estão dispostos no Sistema e-MEC, o qual é mantido pelas instituições (INEP, 2021).

De acordo com os dados do INEP, em 1996, o Brasil tinha 57 Instituições Federais de Ensino Superior (IFES) que totalizavam 1.581 cursos. Nos últimos dados disponibilizados de 2021, o país contava com 110 IFES e 6.669 cursos (INEP, 2021).

2.3 INTELIGÊNCIA COMPUTACIONAL

Na presente seção, apresenta-se a exposição de dados indispensáveis à compreensão do estudo, tais como *aprendizado de máquina* e a hierarquia do aprendizado.

2.3.1 Princípios e capacidades

A base da Inteligência Computacional é sustentada por princípios essenciais que abrangem a adaptação, a aprendizagem, a evolução, a discussão e a imprecisão. A adaptação, em particular, desempenha um papel central, referindo-se à notável capacidade dos sistemas computacionais de ajustarem-se e adaptarem-se

a novas circunstâncias e dados (Bellman, 1978).

Essa capacidade é alcançada por meio da aplicação de algoritmos que possibilitam ao sistema aprender e modificar seu comportamento com base nas informações recebidas.

As técnicas de inteligência computacional utilizadas nas resoluções de problemas apresentam-se com alto grau de eficiência na resolução de atividades envolvendo análise e mineração de dados históricos armazenadas em grandes bases (Bellman, 1978).

A definição dos contextos que envolvem a mineração de dados não é uma tarefa simples. A geração de métodos altamente precisos e rigorosos torna-se impraticável, uma vez que a aplicação da inteligência computacional frequentemente se depara com desafios do mundo real (Abonyi; Feil; Abraham, 2005).

Algumas técnicas utilizadas na mineração de dados com a aplicação de inteligência computacional são capazes de produzir um raciocínio aproximado as respostas que precisamos para a realização do processo de tomada de decisão.

De acordo com o paradigma de aprendizado, as técnicas podem ser classificadas como: aprendizado supervisionado (árvore de decisão*, Naive Baynes e k-nearest neighbor classification*) e aprendizado não supervisionado*: k-means* e em (*expectation maximization*), (Charniak; Mcdermott, 1985).

A definição dos contextos que envolvem a mineração de dados não é uma tarefa simples. A geração de métodos altamente precisos e rigorosos torna-se impraticável, uma vez que a aplicação da inteligência computacional frequentemente se depara com desafios do mundo real (Abonyi; Feil; Abraham, 2005).

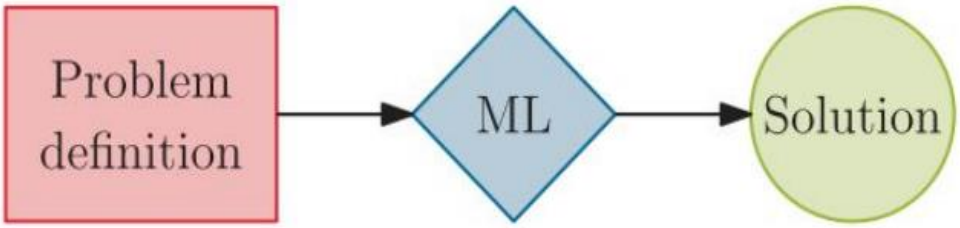
2.3.2 Aprendizado de máquina

Se entende como Aprendizado de Máquina, como uma área de inteligência artificial (IA), que tem a finalidade de desenvolver técnicas computacionais sobre o aprendizado e sistemas que tem a capacidade de adquirir conhecimentos que estão escondidos nos dados de maneira automática (Amorim; Barone; Mansur, 2008).

Para que esta tarefa ocorra de maneira efetiva, é necessário que o sistema de aprendizado realize inferências para que ele possa tomar decisões se baseando em diversas experiências que através das inferências foram acumuladas como base de dados para que a decisão e solução seja bem-sucedida através de análises

anteriores (Mitchell, 1997).

Figura 1 - O aprendizado de máquina atua sozinho para fornecer uma solução para o problema



Fonte: Monard e Baranauskas (2003).

Os diversos sistemas de aprendizado de máquina possuem características comuns e particulares que dá a possibilidade de classificação quanto à linguagem de descrição, forma de aprendizado, paradigma, modo utilizado (Amorim; Barone; Mansur, 2008).

2.3.3 Aprendizado de máquina - Hierarquia do aprendizado

Dentro do universo e do processo de aprendizado de máquina podemos destacar os formatos de aprendizado, que são: aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado e aprendizado por reforço (Amorim; Barone; Mansur, 2008). Nos tópicos abaixo será abordado características fundamentas destas técnicas.

"A hierarquia do aprendizado em Aprendizado de máquina compreende sistemas orientados ao conhecimento que buscam criar estruturas simbólicas compreensíveis para os seres humanos" (Mitchell, 1997, p. 22). Esses sistemas permitem que as máquinas aprendam a partir de dados e experiências, adquirindo conhecimento e melhorando seu desempenho ao longo do tempo.

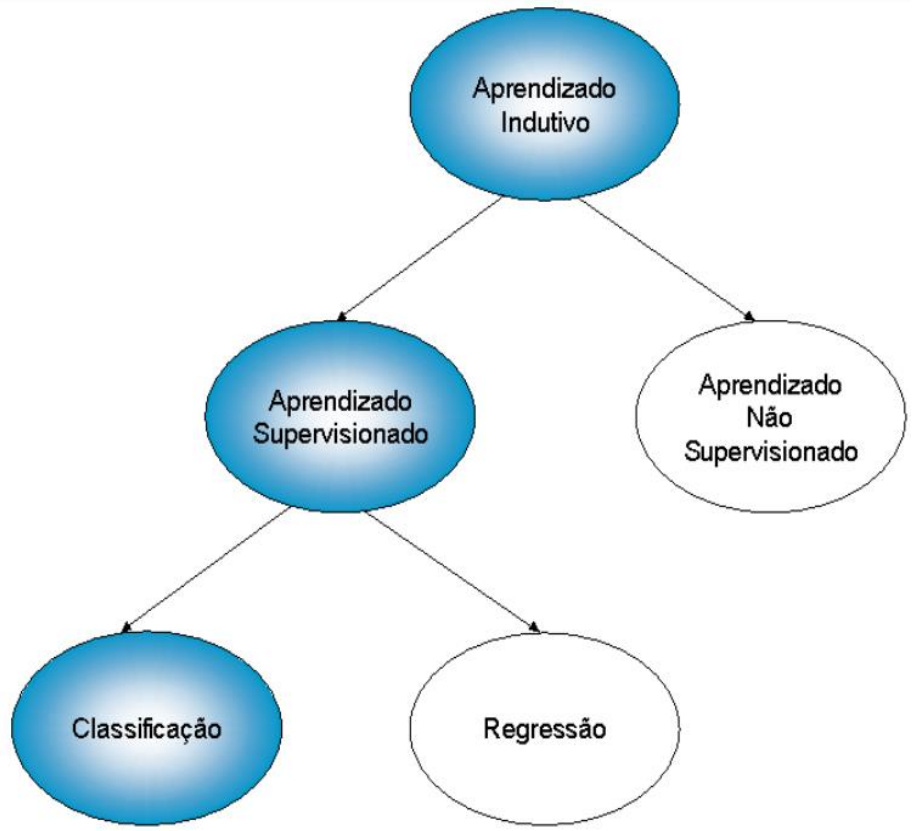
No nível mais baixo dessa hierarquia conforme figura 5, encontramos algoritmos de aprendizado supervisionado, onde a máquina é treinada com exemplos rotulados para realizar previsões ou classificações. Em seguida, temos os algoritmos de aprendizado não supervisionado, que exploram a estrutura dos dados sem rótulos para descobrir padrões e realizar agrupamentos (Amorim; Barone; Mansur, 2008).

No nível seguinte, estão os algoritmos de aprendizado por reforço, que

aprendem através de tentativa e erro, recebendo feedbacks em forma de recompensas ou penalidades. Eles são capazes de tomar decisões otimizadas em ambientes complexos e dinâmicos.

E no topo da hierarquia, estão os sistemas de aprendizado simbólico, que visam criar estruturas compreensíveis para os humanos. Eles utilizam representações simbólicas, como lógica e linguagem natural, para expressar o conhecimento adquirido pela máquina, chamado de aprendizado indutivo. Esses sistemas têm o objetivo de gerar explicações claras e interpretações compreensíveis, facilitando a colaboração entre humanos e máquinas, conforme observa-se na Figura 5:

Figura 2 - Hierarquia do aprendizado



Fonte: Monard e Baranauskas (2003).

A hierarquia do aprendizado em *Aprendizado de máquina* abrange desde algoritmos que aprendem a partir de exemplos e padrões até sistemas que buscam criar estruturas simbólicas compreensíveis. Essa progressão na complexidade do aprendizado permite que as máquinas adquiram conhecimento e interajam de forma mais intuitiva e eficaz com os seres humanos.

2.3.4 Características do Aprendizado Supervisionado

O Aprendizado Supervisionado consiste na técnica de *Aprendizado de máquina* onde os dados são fornecidos diretamente ao algoritmo de treinamento, este por sua vez são chamados de dados rotulados, pois, eles contêm informações sobre o cenário em que foram gerados e o resultado correspondente.

Isso permite que o algoritmo identifique padrões e correlações entre os cenários e os resultados, possibilitando a previsão de saídas únicas para situações similares, conforme preconiza Mitchell (1997):

> Diz-se que um programa de computador aprende com a experiência, E no que diz respeito a alguma classe de tarefas T e medida de desempenho P, se seu desempenho nas tarefas em T, conforme medido por P, melhora com a experiência E (Mitchell, 1997, p. 34).

O Aprendizado Supervisionado pode ser dividido em duas frentes, ou tipos de problemas (MITCHELL, 1997), são eles elencados e devidamente referenciados no Quadro 1, a seguir:

Quadro 1 - Divisões do Aprendizado Supervisionado

Nomenclatura	Classificação
Regressão	Na abordagem de regressão, ocorre a "estimação" dos valores de cada atributo, seguida da identificação de correlações entre esses valores. Em problemas de regressão, um cenário e um objetivo são fornecidos, permitindo que o algoritmo estabeleça relações entre os atributos e aplique a melhor solução possível para atingir o objetivo. Um exemplo comum de aplicação de algoritmos de aprendizado de máquina na regressão é a previsão do comportamento de ações (Delen, 2010).
Classificação	Aqui, os algoritmos são utilizados com a finalidade de solucionar problemas que envolvem a separação de elementos em categorias específicas, como, por exemplo, distinguir fotografias de gatos e cachorros. No entanto, a aplicabilidade desses algoritmos vai além desse contexto, podendo ser ampliada ao incorporar atributos relacionados ao clima, temperatura e previsão do tempo, a fim de determinar se determinado dia é favorável para a prática de atividades esportivas (Fernández; Gil; Mora, 2019).

Conclui-se que o Aprendizado Supervisionado, apresenta uma ampla gama de aplicações e pode ser dividido em duas frentes principais, conforme proposto por

Mitchell (1997). Essas divisões, referenciadas no Quadro 1, fornecem uma estrutura para a classificação dos problemas associados ao Aprendizado Supervisionado.

2.3.5 Características do Aprendizado Não-Supervisionado

Aprendizado não supervisionado, nesta técnica utiliza-se o Aprendizado de máquina para analisar clusters de dados não rotulados, ou seja, aqui os algoritmos aplicados, descobrem padrões escondidos nos dados sem necessidade de supervisão humana, ou seja, o aprendizado não supervisionado é uma abordagem de Aprendizado de máquina que se baseia na análise de dados sem a necessidade de rótulos ou supervisão prévia e o objetivo é encontrar padrões e estruturas ocultas nos dados, permitindo a descoberta de informações relevantes e *insights* sem a necessidade de um conjunto de dados rotulados.

Uma das principais características do aprendizado não supervisionado é a capacidade de identificar agrupamentos ou *clusters* nos dados, conforme descritas detalhadamente no Quadro 2 sobre técnicas utilizadas, logo abaixo.

Isso significa que o algoritmo é capaz de agrupar instâncias de dados similares em grupos, mesmo sem ter informações prévias sobre esses grupos (Mitchell, 1997). Essa capacidade é extremamente útil em tarefas como segmentação de mercado, análise de redes sociais e detecção de fraudes.

O aprendizado não supervisionado é uma abordagem poderosa para explorar grandes volumes de dados sem a necessidade de rótulos prévios. Suas principais características incluem a identificação de agrupamentos, a redução de dimensionalidade e a detecção de anomalias.

Apesar dos desafios inerentes, essa abordagem tem se mostrado fundamental para a compreensão e utilização dos dados em diversas áreas de aplicação.

Quadro 2 - Técnicas Utilizadas

Nomenclatura	Características
Redução de dimensionalidade	Essa técnica é aplicada quando o número de atributos ou dimensões dos dados é elevado. A finalidade é reduzir a quantidade de entradas, mantendo apenas as mais relevantes, otimizando, assim, o algoritmo, ou seja, permite encontrar representações mais compactas dos dados, reduzindo a quantidade de variáveis e mantendo a essência das informações originais. A redução de dimensionalidade pode ser empregada no pré-processamento de dados e no aprimoramento de imagens. Isso facilita a visualização e interpretação dos dados, além de contribuir para a eficiência dos algoritmos de aprendizado de máquina (Chui; Fung; Lok, 2020).
Clusterização	Trata-se de uma técnica que tem por objetivo agrupar dados não rotulados com base em suas similaridades ou diferenças. Essa abordagem revela-se bastante útil em aplicações como compressão de imagens, criação automática de álbuns, detecção de imagens, entre outras (Fernández; Gil; Mora, 2019).
Associação	Por sua vez, refere-se a uma técnica que utiliza diferentes regras para encontrar correlações entre variáveis. É comumente empregada em análises e recomendações, como no caso de recomendar produtos a partir da seguinte lógica: "pessoas que compraram X também compraram Y" (Delen, 2010).

Além disso, o aprendizado não supervisionado também é utilizado para detecção de anomalias. Os algoritmos são capazes de identificar padrões atípicos nos dados, indicando a presença de comportamentos incomuns ou suspeitos. Isso tem aplicações em áreas como segurança cibernética, monitoramento de sistemas e detecção de fraudes.

2.3.6 Características do Aprendizado Semi-Supervisionado

O aprendizado semi-supervisionado é capaz de lidar com grandes conjuntos de dados, onde apenas uma pequena fração é rotulada, tornando-o vantajoso em muitos cenários do mundo real.

De acordo com o que preconiza Mitchell (1997), é uma abordagem que combina dados rotulados e não rotulados para treinar um método preditivo (Adejo; Connolly, 2017).

Nesse contexto, características específicas são observadas. Além disso, ele explora a estrutura e a distribuição dos dados não rotulados para melhorar o desempenho do método. O aprendizado semi-supervisionado pode ser aplicado em diversas áreas, como classificação de documentos, reconhecimento de padrões e processamento de imagens, entre outros.

2.3.7 Características do Aprendizado por Reforço

O Aprendizado de Máquina por Reforço é uma abordagem que permite que um agente aprenda a tomar decisões por meio da interação com um ambiente (Mitchel,1997).

O aprendizado por reforço envolve a noção de exploração e explotação, onde o agente busca equilibrar a obtenção de recompensas imediatas e a busca por ações que levem a maiores recompensas a longo prazo.

Outra característica importante é a presença de um processo de aprendizagem sequencial, onde as ações tomadas em um momento influenciam as futuras ações e decisões. Adicionalmente, o aprendizado de máquina por reforço requer a definição de uma função de recompensa que guie o agente na busca pela melhor política de ação (Russell; Norvig, 2013).

2.4 APRENDIZADO DE MÁQUINA E OS ALGORITMOS DE CLASSIFICAÇÃO

A classificação em aprendizagem de máquina é um procedimento fundamental para identificar características e rotular as saídas correspondentes dentro de um contexto específico (Mitchel,1997).

Essa técnica é amplamente utilizada em diversos domínios, como reconhecimento de padrões, análise de dados e tomada de decisões automatizadas. Nesse processo, algoritmos de aprendizagem de máquina são treinados com um conjunto de dados previamente rotulados, a fim de criar um método capaz de classificar novas instâncias com base em suas características (Adejo; Connolly, 2017).

A classificação é geralmente realizada utilizando-se técnicas estatísticas e algoritmos de aprendizagem supervisionada, como árvores de decisão, redes neurais e máquinas de vetor de suporte (Mitchel,1997).

O objetivo principal da classificação é automatizar o processo de atribuir rótulos a dados não rotulados, permitindo a organização e a compreensão de grandes volumes de informações de forma eficiente (Adejo; Connolly, 2017).

Esses métodos permitem que um sistema automatizado aprenda a reconhecer padrões e tome decisões com base em dados previamente fornecidos. Os algoritmos de classificação, em particular, são capazes de agrupar dados em categorias ou classes distintas, tornando-se uma ferramenta valiosa para problemas de classificação.

Além disso, destacamos a importância do pré-processamento dos dados, da seleção adequada de características e da validação dos resultados obtidos. O uso de *Aprendizado de máquina* e algoritmos de classificação apresenta um grande potencial para auxiliar na tomada de decisões em diferentes áreas, como medicina, finanças, marketing e reconhecimento de padrões em imagens, evasão escolar (Adejo; Connolly, 2017).

2.4.1 Classificação

A classificação, segundo o que preconiza Mitchell (1997), é um processo que visa prever a categoria de uma observação específica, o objetivo é desenvolver um "classificador" capaz de gerar uma classificação qualitativa para uma nova observação com base em dados de entrada, os quais incluem observações já definidas e classificadas.

Como exemplo, podemos considerar um classificador que utilize informações não observadas de um paciente e o classifique como doente ou não-doente. Essa abordagem pode ser aplicada em diversas áreas, como medicina, ciência de dados e aprendizado de máquina.

O desenvolvimento de classificadores eficientes requer a utilização de algoritmos e técnicas adequadas, além de um conjunto de dados bem definido e validado para treinamento e teste.

A classificação é uma importante ferramenta para auxiliar na tomada de decisões e na identificação de padrões e tendências em dados não observados (MITCHELL, 1997).

Existem diversos algoritmos classificadores que podem ser utilizados em uma pesquisa para resolver problemas de aprendizado de máquina, porém cada situação

é específica e cada algoritmo poderá entregar resultados diferentes.

Seguem alguns exemplos de algoritmos classificadores utilizados com a escolha do algoritmo mais adequado depende do contexto específico do problema, do conjunto de dados disponíveis e dos objetivos da classificação, veja os principais conceitos dos algoritmos possíveis no Quadro 3, a seguir:

Quadro 3 - Algoritmos Possíveis

Algoritmo	Classificação
Naive Bayes	É um algoritmo baseado no teorema de *Bayes*, que assume independência condicional entre os atributos. Ele é amplamente utilizado em problemas de classificação de texto, como a detecção de spam em e-mails (Charniak; Mcdermott, 1985).
Árvores de Decisão	Esse algoritmo constrói uma estrutura de árvore na qual cada nó interno representa um teste em um atributo, e as folhas representam as classes de saída. É uma abordagem intuitiva e fácil de interpretar (Chui; Fung; Lok, 2020).
Random Forest	Trata-se de um conjunto de árvores de decisão, onde cada árvore é construída de forma aleatória. O resultado é obtido por meio de uma votação das previsões de todas as árvores (Gamie; El-Seoud; Salama, 2020).
Support Vector Machines (SVM)	Esse algoritmo busca encontrar um hiperplano que maximize a margem de separação entre as classes. É eficiente em problemas de classificação binária, mas também pode ser estendido para classificação multi-classe (Géron, 2019).
Redes Neurais Artificiais	São modelos inspirados no funcionamento do cérebro humano, compostos por várias camadas de neurônios interconectados. Podem ser aplicadas em problemas de classificação complexos e requerem um grande volume de dados para treinamento (Delen, 2010).
k-Nearest Neighbors (k-NN)	Esse algoritmo classifica uma observação com base nas classes de seus k vizinhos mais próximos. Ele utiliza a medida de distância para determinar a proximidade entre as observações (Géron, 2019).
Logistic Regression (Regressão Logística)	Apesar do nome, a regressão logística é um algoritmo de classificação que estima a probabilidade de pertencer a uma determinada classe. É comumente usados em problemas de classificação binária (Delen, 2010).
Gradient Boosting	Essa é uma técnica que combina vários modelos de classificação mais fracos em um modelo forte. O algoritmo gera previsões iniciais e, em seguida, ajusta subsequentemente os modelos para corrigir os erros anteriores (Grus, 2016).

Algoritmo	Classificação
Máquinas de Vetores de Suporte com Kernel (Kernel SVM)	É uma extensão do algoritmo SVM que utiliza funções de kernel para mapear os dados para um espaço de maior dimensionalidade. Isso permite que ele lide com problemas de classificação não linear (Fernández; Gil; Mora, 2019).
Redes Neurais Convolucionais (CNN)	São redes neurais especializadas para processar dados estruturados em formato de grade, como imagens. Elas são muito eficazes em problemas de classificação de imagens e visão computacional (Grus, 2016).
Gaussian Naive Bayes	É uma variação do *Naive Bayes* que assume que os dados seguem uma distribuição normal (gaussiana). É especialmente útil quando os atributos são contínuos (Charniak; Mcdermott, 1985).
Árvores de Decisão Estocásticas (Randomized *Decision Trees*)	São variantes das árvores de decisão tradicionais, nas quais a seleção dos atributos e a divisão dos nós são realizadas de forma aleatória. Essa aleatoriedade ajuda a evitar o sobre ajuste e melhora a robustez do modelo (Gamie; El-Seoud; Salama, 2020).
AdaBoost	É um algoritmo de *boosting* que combina várias instâncias de um classificador fraco para formar um classificador forte. Ele atribui maior peso às observações classificadas incorretamente nas iterações anteriores, a fim de melhorar a precisão do modelo (Grus, 2016).
Redes Bayesianas	São modelos probabilísticos que representam as relações de dependência entre os atributos por meio de um grafo acíclico direcionado. Esses modelos são úteis quando há incerteza sobre as relações entre os atributos (Charniak; Mcdermott, 1985).
Regressão Logística Multinomial	É uma extensão da regressão logística para problemas de classificação com mais de duas classes. Ele estima as probabilidades de pertencer a cada classe e seleciona a classe com maior probabilidade (Fernández; Gil; Mora, 2019).

Portanto, a escolha de um algoritmo classificador adequado é essencial para o sucesso de um projeto de aprendizado de máquina. Cada algoritmo possui suas próprias características e suposições subjacentes, o que pode influenciar significativamente os resultados obtidos.

É importante considerar cuidadosamente o contexto específico do problema, o conjunto de dados disponíveis e os objetivos da classificação ao selecionar o algoritmo mais adequado.

Com a aplicação correta dos algoritmos classificadores, é possível obter insights valiosos, tomar decisões informadas e identificar padrões e tendências

ocultos nos dados não observados.

A classificação, portanto, se mostra uma poderosa ferramenta no campo da ciência de dados e do aprendizado de máquina, impulsionando avanços e descobertas em diversas áreas de pesquisa.

2.4.2 Regressão

Se trata de uma técnica estatística amplamente utilizada para modelar e prever relações entre variáveis pois, segundo o que preconizam Russell e Norvig (2013), a regressão é um algoritmo estatístico que estima os parâmetros de um modelo matemático com base em dados históricos.

Esse modelo descreve a relação entre uma variável dependente e uma ou mais variáveis independentes. Através dessa abordagem, é possível obter insights valiosos sobre fenômenos complexos e realizar previsões confiáveis em diversas áreas, como economia, ciências sociais, medicina e engenharia.

De acordo Mitchell (1997), a regressão é uma poderosa ferramenta estatística que permite modelar e prever relações entre variáveis. O objetivo principal desse algoritmo é estimar os parâmetros do modelo matemático que descreve a relação entre a variável dependente e as variáveis independentes. Essa estimativa é feita através de técnicas estatísticas, como o método dos mínimos quadrados.

No processo de aplicação do algoritmo de regressão, é importante seguir alguns passos fundamentais, Mitchell (1997). Primeiramente, ocorre a seleção das variáveis independentes que serão incluídas no modelo. Essa seleção pode ser baseada em conhecimento prévio sobre o problema em estudo ou utilizando técnicas estatísticas, como a análise de correlação.

Uma vez selecionadas as variáveis, o próximo passo consiste em ajustar o modelo de regressão aos dados. Esse ajuste envolve a estimativa dos coeficientes do modelo por meio de técnicas estatísticas, como o método dos mínimos quadrados, Russell e Norvig (2013).

O objetivo é encontrar os valores dos coeficientes que melhor se ajustem aos dados observados. Após o ajuste do método, é essencial avaliar a qualidade desse ajuste, o que pode ser realizado utilizando métricas estatísticas, como o coeficiente de correlação (R^2) (Russell; Norvig, 2013).

Além disso, é importante realizar uma análise dos resíduos, que são as

diferenças entre os valores observados e os valores previstos pelo modelo, Mitchell (1997). Essa análise dos resíduos permite verificar se o modelo captura adequadamente a estrutura dos dados, garantindo assim resultados confiáveis.

O algoritmo de regressão é uma técnica estatística essencial para a modelagem e previsão de relações entre variáveis. Sua aplicação oferece suporte à tomada de decisões em diversas áreas do conhecimento, porém requer uma cuidadosa seleção de variáveis, avaliação do ajuste do modelo e interpretação correta dos resultados, (Bengio, 2009).

Entre os diversos algoritmos de regressão para aplicação em *Aprendizado de máquina*, pode-se apresentar os mais importantes dentro do universo da deste trabalho de pesquisa, que são:

Quadro 4 - Algoritmos de Regressão em Escolha

Nomenclatura	**Características**
Regressão Linear Simples	É um algoritmo de regressão que estabelece uma relação linear entre uma variável dependente e uma única variável independente. Por exemplo, pode ser usado para prever o preço de uma casa com base na sua área (Chui; Fung; Lok, 2020).
Regressão Linear Múltipla	Similar à regressão linear simples, porém permite modelar a relação entre uma variável dependente e múltiplas variáveis independentes. Por exemplo, pode ser usado para prever o salário de um indivíduo com base em variáveis como idade, educação e experiência (Chui; Fung; Lok, 2020).
Regressão Logística	É um algoritmo utilizado para modelar a relação entre uma variável dependente binária (0 ou 1) e um conjunto de variáveis independentes. É frequentemente utilizado em problemas de classificação, como prever se um paciente tem uma determinada doença com base em seus sintomas (Delen, 2010).
Regressão Polinomial	Este algoritmo permite modelar uma relação não-linear entre a variável dependente e as variáveis independentes. Pode ser utilizado quando a relação entre as variáveis não é estritamente linear. Por exemplo, pode ser usado para modelar a taxa de crescimento de uma população ao longo do tempo (Gamie; El-Seoud; Salama, 2020).
Regressão Ridge e Lasso	São técnicas de regressão utilizadas quando existe multicolinearidade (alta correlação) entre as variáveis independentes. Elas aplicam penalidades aos coeficientes do modelo, ajudando a evitar *overfitting* e melhorando a estabilidade dos resultados (Fernández; Gil; Mora, 2019).

Nomenclatura	Características
Regressão de Árvore de Decisão	Este algoritmo utiliza uma estrutura de árvore para modelar a relação entre as variáveis dependentes e independentes. Ele divide os dados em segmentos com base em determinados critérios, construindo uma árvore que permite fazer previsões (Gamie; El-Seoud; Salama, 2020).
Regressão de Floresta Aleatória	É uma extensão da regressão de árvore de decisão que utiliza várias árvores de decisão para realizar a previsão. Cada árvore é construída em uma amostra aleatória dos dados e, em seguida, as previsões de todas as árvores são combinadas para obter uma previsão final mais precisa (Fernández; Gil; Mora, 2019).
Regressão de Máquina de Vetores de Suporte (SVM)	Este algoritmo é usado tanto para classificação quanto para regressão. Na regressão, ele mapeia os dados para um espaço de alta dimensão e encontra um hiperplano ótimo para prever a variável dependente (Delen, 2010).
Regressão Bayesiana	Este algoritmo utiliza o teorema de *Bayes* para realizar a regressão. Ele estima a distribuição de probabilidade dos parâmetros do modelo com base em dados observados e fornece uma previsão probabilística para a variável dependente (Charniak; Mcdermott, 1985).
Regressão por Redes Neurais	As redes neurais são modelos de aprendizado de máquina inspirados no funcionamento do cérebro. Na regressão, elas podem ser usadas para modelar relações complexas entre variáveis dependentes e independentes, permitindo a captura de padrões não lineares (Delen, 2010).

A regressão desempenha um papel fundamental na análise de dados e na compreensão de relações entre variáveis. Ao aplicar essa técnica estatística, é possível obter informações que auxiliam na tomada de decisões em diversas áreas do conhecimento.

2.4.3 Agrupamento

Os algoritmos de agrupamento são baseados em diferentes abordagens, como o método de partição, hierárquico, baseado em densidade e baseado em grade. Cada abordagem tem suas vantagens e desvantagens, e a escolha do algoritmo adequado depende das características dos dados e dos objetivos da análise.

Além disso, é importante considerar medidas de avaliação, como a validação interna e externa dos agrupamentos obtidos, a fim de verificar sua qualidade e

robustez (Mitchell, 1997).

A implementação dos algoritmos de agrupamento geralmente envolve a definição de parâmetros, como o número de clusters desejado e a função de similaridade utilizada para medir a proximidade entre os dados.

Conforme Mitchell (1997), além disso, é fundamental realizar a pré-processamento dos dados, incluindo a remoção de ruídos e a normalização das variáveis, a fim de garantir resultados mais precisos.

Segue abaixo alguns algoritmos utilizados em *Aprendizado de máquina*:

Quadro 5 - Algoritmos Utilizados em *Aprendizado de máquina*

Nomenclatura	Classificação
K-means	É um dos algoritmos mais populares e simples. Ele divide os dados em k *clusters,* onde k é um número pré-definido, minimizando a soma dos quadrados das distâncias entre os pontos e seus respectivos centroides (Chui; Fung; Lok, 2020).
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	É um algoritmo baseado em densidade que agrupa os pontos com base na densidade dos vizinhos. Ele é capaz de identificar clusters de diferentes formas e tamanhos, além de detectar pontos de ruído (Delen, 2010).
Hierarchical Clustering (Agrupamento Hierárquico)	É um algoritmo que cria uma estrutura hierárquica de clusters, formando uma árvore de divisão (dendrograma). Pode ser aglomerativo (começando com cada ponto como um cluster separado e mesclando-os) ou divisivo (começando com todos os pontos em um único cluster e dividindo-os) (Gamie; El-Seoud; Salama, 2020).
Gaussian Mixture Models (GMM)	É um algoritmo que modela os dados como uma mistura de distribuições gaussianas. Ele assume que os dados foram gerados por um conjunto de distribuições gaussianas não observadas e estima os parâmetros dessas distribuições para realizar o agrupamento (Géron, 2019).
Mean Shift	É um algoritmo que encontra os máximos locais da função de densidade dos dados, deslocando iterativamente os pontos em direção aos máximos até convergir. Os pontos que convergem para o mesmo máximo local são agrupados juntos (Gamie; El-Seoud; Salama, 2020).
Agglomerative Clustering (Agrupamento Aglomerativo)	É um algoritmo hierárquico que começa com cada ponto como um cluster separado e, em seguida, mescla iterativamente os clusters com base em uma medida de proximidade até obter os clusters finais (Géron, 2019).

Nomenclatura	Classificação
Spectral Clustering (Agrupamento Espectral)	*Spectral Clustering* (Agrupamento Espectral): é um algoritmo que utiliza a matriz de similaridade dos dados para realizar o agrupamento. Ele mapeia os pontos em um espaço de maior dimensão e, em seguida, aplica um algoritmo de agrupamento nesse espaço transformado (Gamie; El-Seoud; Salama, 2020).
BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)	É um algoritmo que constrói uma árvore hierárquica balanceada chamada CF (*Clustering Feature*). Ele utiliza uma estrutura de dados compacta para realizar o agrupamento de forma eficiente em grandes conjuntos de dados (Delen, 2010).
OPTICS (Ordering Points To Identify the Clustering Structure)	É um algoritmo baseado em densidade que gera uma ordenação dos pontos, levando em consideração a densidade e a distância entre eles. Ele permite a detecção de clusters de diferentes densidades e formas (Delen, 2010).
SOM (Self-Organizing Map)	é um tipo especial de rede neural que mapeia os dados de entrada em uma grade bidimensional. Os pontos de dados próximos no mapa são agrupados juntos, permitindo a visualização e análise dos dados em um espaço de menor dimensão (Chui; Fung; Lok, 2020).

Para garantir a eficácia dos algoritmos de agrupamento em Aprendizado de máquina, é essencial considerar uma etapa adicional de pós-processamento dos resultados obtidos.

Nessa fase, é possível realizar a análise dos clusters gerados e interpretar os padrões encontrados, buscando insights relevantes para a tomada de decisão. Além disso, é importante ressaltar que a seleção dos algoritmos de agrupamento adequados deve levar em conta não apenas as características dos dados, mas também as necessidades específicas do problema em questão.

Ao adotar uma abordagem criteriosa na escolha dos métodos e no tratamento dos dados, é possível obter agrupamentos de alta qualidade, impulsionando assim os resultados e a aplicabilidade do Aprendizado de máquina.

2.5 ÁRVORE DE DECISÃO: FUNDAMENTOS A TÉCNICA DE APRENDIZADO DE MÁQUINA

A árvore de decisão é uma técnica fundamental no campo do aprendizado de máquina. Ela é amplamente utilizada para classificar e prever dados com base em

uma sequência de regras de decisão.

Essa abordagem consiste em dividir o conjunto de dados em subconjuntos menores, com base nas características mais relevantes, até que sejam alcançadas decisões finais ou previsões. Nas próximas seções será abordado características e informações sobre a técnica Árvore de Decisão e demais informações relevantes para esta pesquisa.

2.5.1 Árvore de Decisão: historicidade e conceitos

As Árvore de Decisão são técnicas mais populares e amplamente utilizadas em aprendizado de máquina. Essa abordagem é fundamental para a construção de modelos preditivos que possam realizar decisões com base em um conjunto de características fornecidas (Mitchell, 1997). Neste sentido, exploraremos conceitos iniciais relacionados às árvores de decisão, no contexto do aprendizado de máquina.

Árvore de Decisão, é um método eficaz para representar o conhecimento em forma de regras de decisão. De acordo com Russel e Norvig (2013), "as Árvore de Decisão permitem uma estrutura clara e intuitiva para tomar decisões com base em múltiplas variáveis".

As Árvore de Decisão são úteis para mapear relações complexas entre variáveis de entrada e saída. "as Árvore de Decisão podem lidar com problemas de classificação e regressão, fornecendo respostas claras e interpretações compreensíveis" (Russel; Norvig, 2013).

As Árvore de Decisão são construídas a partir de um conjunto de exemplos de treinamento, onde cada exemplo possui características e uma classe associada. "as Árvore de Decisão aprendem a mapear as características para a classe correspondente, permitindo a realização de previsões precisas em novos exemplos" (Mitchell, 1997).

De acordo com Bengio (2009), as Árvore de Decisão têm a capacidade de lidar com atributos não lineares e interações complexas entre variáveis, tornando-se uma escolha viável em problemas de alta dimensionalidade". Existe também importância da interpretabilidade das Árvore de Decisão, permite entender como as decisões são tomadas com base nas características de entrada.

2.5.2 Como funciona a Árvore de Decisão

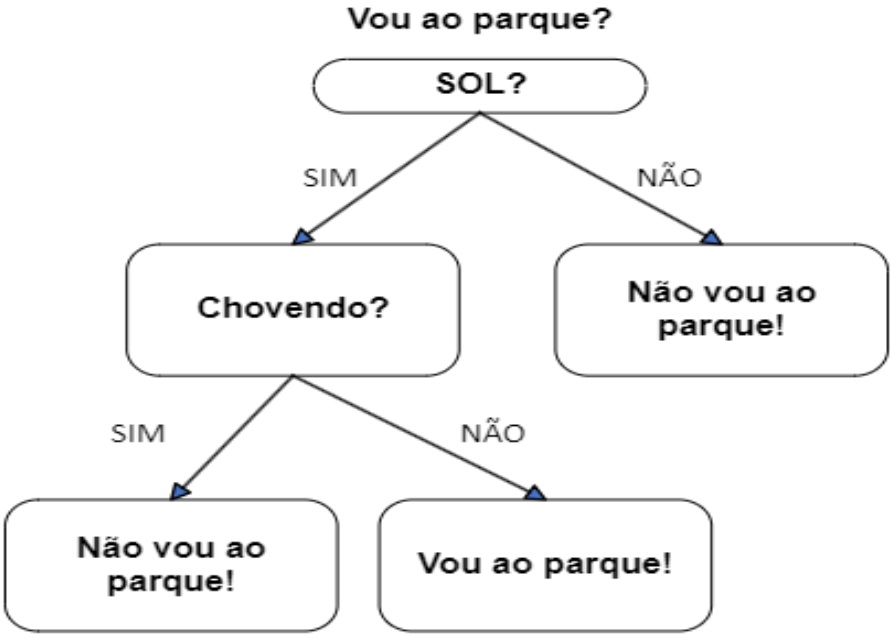
As árvores de decisão são um método eficaz para representar o conhecimento em forma de regras de decisão. Elas permitem uma estrutura clara e intuitiva para tomar decisões com base em múltiplas variáveis (Russel; Norvig, 2013).

Essa abordagem é útil para mapear relações complexas entre variáveis de entrada e saída, sendo capaz de lidar com problemas de classificação e regressão.

No contexto do aprendizado de máquina, as Árvore de Decisão são construídas a partir de um conjunto de exemplos de treinamento, onde cada exemplo possui características e uma classe associada (Mitchell, 1997).

O objetivo é aprender a mapear as características para a classe correspondente, permitindo realizar previsões precisas em novos exemplos conforme exemplificado na Figura 6.

Figura 3 - Metodologia e Itinerário Metodológico



Fonte: do autor (2023).

Além disso, as árvores de decisão têm a capacidade de lidar com atributos não lineares e interações complexas entre variáveis, o que as torna uma escolha viável em problemas de alta dimensionalidade.

Sua interpretabilidade também é um fator importante, pois permite entender

como as decisões são tomadas com base nas características de entrada (Bengio, 2020).

2.5.3 Representação da Árvore de Decisão

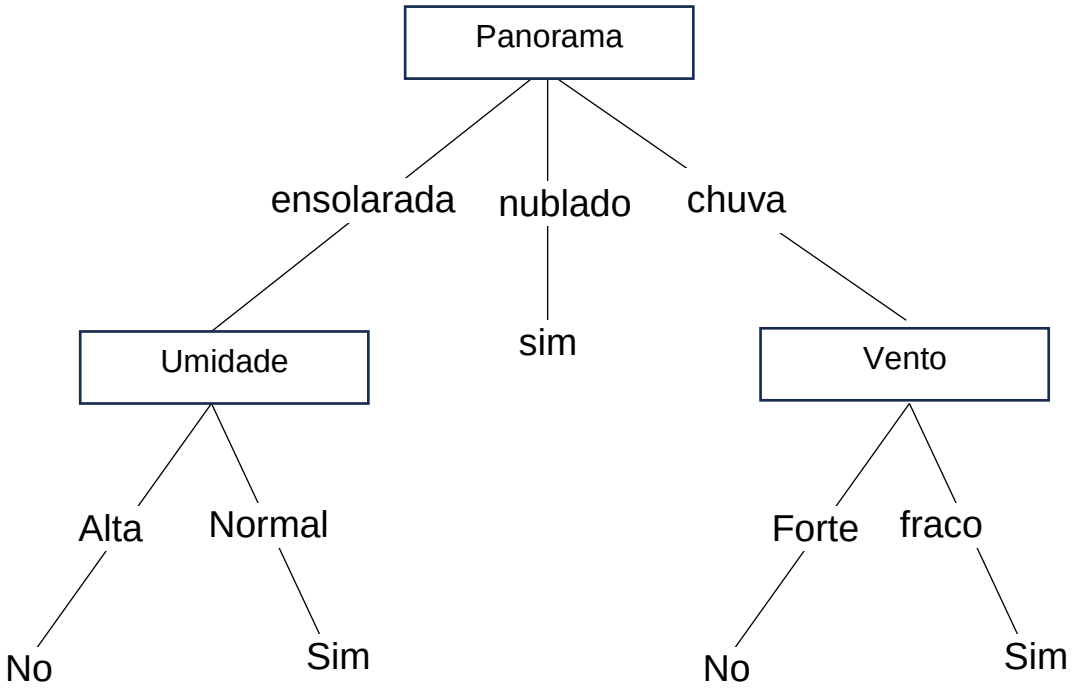
A representação da Árvore de Decisão é uma forma gráfica que ilustra as etapas do processo de tomada de decisão. A estrutura da árvore é composta por nós, que representam as escolhas a serem feitas, e ramos, que representam as consequências dessas escolhas.

2.5.4 Conceitos e representação

A aprendizagem por Árvore de Decisão é um método usado para aproximar funções-alvo discretas, em que a função aprendida é representada por uma árvore de decisão e as árvores aprendidas também podem ser transformadas em conjuntos de regras "se-então" para facilitar a compreensão humana, tal qual preconiza Mitchel (1997).

Esses métodos de aprendizagem são bastante populares entre os algoritmos de inferência indutiva e têm sido amplamente aplicados em diversas tarefas, desde o diagnóstico de casos médicos até a avaliação do risco de crédito de candidatos a empréstimos conforme observa-se na Figura 9 a seguir:

Figura 4 - Representação da Árvore de Decisão - *Decision Tree*



Fonte: *Machine Learning,* Mitchel (1997).

Dessa forma essa árvore bem representa uma análise de possibilidades de se jogar ou não o ténis em uma manhã específica.

2.5.5 Exemplo: decisão para um eventual jogo de ténis

Ao analisar-se essa estrutura se observa que a o panorama inicial está nas possibilidades de acontecer algo em função da possibilidade de jogar ou não ténis ao tempo em que em cada um dos lados (direita e esquerda), conforme os dados disponíveis na estrutura.

Neste exemplo, é possível analisar se o dia está apto ao "Jogo de Tênis", por exemplo e nesse sentido verifica-se que ele é classificado ao percorrê-la até o nó folha adequada, retornando então a classificação associada a essa folha (neste caso, "Sim" ou "Não") que pode ser extraída de acordo com os dados e essa árvore acaba por classificar as manhãs de sábado se está apta para jogar tênis, em todos os cenários é possível coletar dados de possibilidades positivas e negativas.

No caso de uma manhã ensolarada ainda é possível a verificação de humidade para se tornar uma decisão favorável, em outro caso, se nublado, é favorável a prática do exporte e se estiver com chuva fraca também é possível a

prática, para poder chegar a esta decisão as Árvores de Decisão classificam instâncias.

Ao percorrê-las da raiz até um nó folha, fornecendo a classificação correspondente onde cada nó da árvore realiza um teste em algum atributo da instância, e cada ramo representa um resultado possível desse teste (Mitchel, 1997).

2.5.6 Árvore de Decisão como método de referências

A aplicação da Árvore de Decisão na análise de dados e na criação de método de dados universitários com o objetivo de identificar a evasão escolar tem se mostrado uma abordagem promissora.

A evasão escolar é uma preocupação recorrente nas instituições de ensino superior, e compreender os fatores que levam os estudantes a abandonarem os cursos é essencial para desenvolver estratégias de prevenção e intervenção adequadas.

A utilização da Árvore de Decisão permite a construção de um método que analisa características dos estudantes, como idade, gênero, desempenho acadêmico, situação socioeconômica e histórico escolar, para identificar padrões e tomar decisões relacionadas à probabilidade de evasão.

Esse método pode ser treinado com dados históricos de estudantes que evadiram ou que permaneceram na universidade, criando um sistema de previsão capaz de auxiliar na identificação precoce de alunos com maior risco de abandonar os estudos.

Ao analisar os resultados obtidos com a aplicação da Árvore de Decisão, é possível identificar os fatores de maior relevância na predição da evasão escolar (Adejo; Connolly, 2017).

Isso permite que as instituições de ensino adotem medidas preventivas direcionadas, como oferecer suporte acadêmico adicional, programas de orientação e aconselhamento, bolsas de estudo ou atividades extracurriculares, com base nas necessidades específicas de cada aluno.

É importante ressaltar que a aplicação da Árvore de Decisão na análise de evasão escolar requer a disponibilidade de um conjunto de dados representativo e confiável, com informações relevantes sobre os estudantes e seu progresso acadêmico.

Além disso, é necessário realizar uma avaliação contínua do método e atualizá-lo regularmente, à medida que novos dados se tornam disponíveis, a fim de garantir sua precisão e eficácia (Russell; Norvig, 2013).

A aplicação da Árvore de Decisão na análise de dados universitários para identificar a evasão escolar permite uma abordagem mais proativa na gestão acadêmica, possibilitando a implementação de ações preventivas e estratégicas para aumentar a retenção dos estudantes (Mitchell, 1997).

A utilização dessa abordagem pode contribuir significativamente para a melhoria dos índices de conclusão de cursos e para o desenvolvimento de uma educação de qualidade, promovendo o sucesso e o bem-estar dos estudantes universitários.

2.5.7 Árvore de Decisão: relevâncias, vantagens e desvantagens

As árvores de decisão são uma técnica popular de aprendizado de máquina que permite a tomada de decisões com base em estruturas hierárquicas. Neste trabalho, exploraremos as relevâncias, vantagens e desvantagens dessa abordagem. As árvores de decisão oferecem uma compreensão clara e intuitiva dos dados, tornando-as valiosas para análise e previsão.

No entanto, sua tendência a overfitting e a necessidade de pré-processamento adequado são considerações importantes a serem feitas. Vamos examinar mais de perto esses aspectos e como eles impactam o uso das árvores de decisão em problemas de aprendizado de máquina.

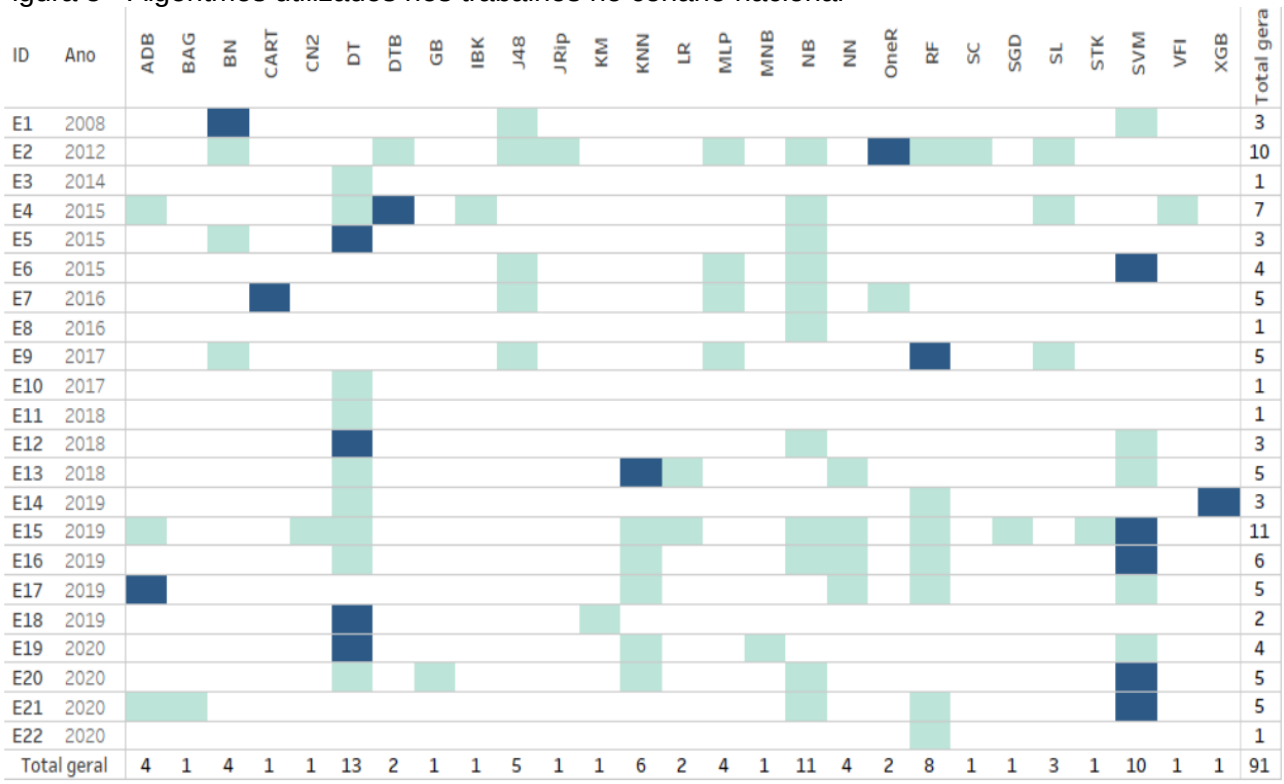
2.5.8 Relevância de uso

Após uma análise abrangente dos principais algoritmos de agrupamento no contexto do *Aprendizado de máquina*, torna-se pertinente a compreensão dos fundamentos que sustentam a seleção da estrutura de classificação conhecida como *Decision Tree* (Árvore de Decisão) e os motivos que justificam sua escolha como base orientadora desta pesquisa.

A relevância dessa abordagem reside no fato de que a Árvore de Decisão é um método amplamente utilizado e comprovado para realizar tarefas de classificação e regressão em diversas áreas de estudo.

Conforme preconiza Silva, Almeida e Ramalho (2020), o Árvore de Decisão aprece em primeiro lugar no cenário nacional como algoritmo mais utilizado na análise de dados, pois ele aparece com um número total de 13 trabalhos que foram pesquisados no total de 22, além disso é o algoritmo que apresenta com melhor desempenho em pelo menos 4 trabalhos no cenário pesquisado, conforme visto na imagem a seguir.

Figura 5 - Algoritmos utilizados nos trabalhos no cenário nacional



ID	Ano	ADB	BAG	BN	CART	CN2	DT	DTB	GB	IBK	J48	JRip	KM	KNN	LR	MLP	MNB	NB	NN	OneR	RF	SC	SGD	SL	STK	SVM	VFI	XGB	Total gera
E1	2008			■							▒															▒			3
E2	2012			▒				▒			▒	▒				▒		▒		■	▒	▒		▒					10
E3	2014						▒																						1
E4	2015	▒					▒	■		▒								▒						▒			▒		7
E5	2015			▒			■											▒											3
E6	2015										▒					▒		▒								■			4
E7	2016				■						▒					▒		▒		▒									5
E8	2016																	▒											1
E9	2017			▒							▒					▒					■			▒					5
E10	2017						▒																						1
E11	2018						▒																						1
E12	2018						■											▒								▒			3
E13	2018						▒							■	▒				▒							▒			5
E14	2019						▒														▒							■	3
E15	2019	▒				▒	▒							▒	▒			▒	▒		▒		▒		▒	■			11
E16	2019						▒							▒				▒	▒		▒					■			6
E17	2019	■												▒					▒		▒					▒			5
E18	2019						■					▒																	2
E19	2020						■							▒			▒									▒			4
E20	2020						▒		▒					▒				▒								■			5
E21	2020	▒	▒															▒			▒					■			5
E22	2020																				▒								1
Total geral		4	1	4	1	1	13	2	1	1	5	1	1	6	2	4	1	11	4	2	8	1	1	3	1	10	1	1	91

Fonte: Silva, Almeida e Ramalho (2020).

De acordo com Mitchell (1997), as Árvores de Decisão são métodos de aprendizado de máquina que estruturam o conhecimento adquirido a partir de um conjunto de dados para realizar previsões ou tomar decisões.

Essas estruturas são particularmente eficazes em problemas nos quais a relação entre as variáveis de entrada e a variável de saída é não linear e complexa.

A vantagem das Árvores de Decisão está na sua capacidade de representar de forma explícita as decisões tomadas em cada nó e as condições que levam a essas decisões, tornando o processo de interpretação e explicação mais acessível (Russell; Norvig, 2013).

2.5.9 Vantagens e desvantagens

No entanto, é importante reconhecer que as Árvore de Decisão também apresentam algumas limitações e desvantagens. Um dos principais desafios é o potencial de supera juste (*overfitting*), no qual a árvore se ajusta excessivamente aos dados de treinamento, resultando em baixa capacidade de generalização para novos dados.

Além disso, a sensibilidade a pequenas variações nos dados de entrada pode levar a árvores de decisão diferentes, o que pode impactar a estabilidade e consistência dos resultados.

Apesar dessas desvantagens, as Árvores de Decisão possuem várias vantagens. Elas são facilmente interpretáveis, permitindo a compreensão direta das decisões tomadas em cada etapa.

Além disso, a estrutura hierárquica das árvores permite uma organização clara do conhecimento, tornando-as úteis para a exploração e extração de informações relevantes dos dados. Adicionalmente, as Árvores de Decisão podem lidar com dados categóricos e numéricos, e sua construção e utilização são computacionalmente eficientes em comparação com outros algoritmos mais complexos.

A escolha da Árvore de Decisão como estrutura de classificação nesta pesquisa se baseia em sua comprovada eficácia na resolução de problemas de classificação e sua capacidade de representação explícita do conhecimento. Embora apresente desvantagens, as vantagens oferecidas pela utilização da Árvores de Decisão, como sua interpretabilidade e facilidade de implementação, justificam sua seleção como base para a investigação científica em questão.

2.6 TRABALHOS RELACIONADOS

São descritos alguns trabalhos relacionados a relação entre eles e o trabalho proposto, além da relação com os desafios mencionados.

2.6.1 Uso de algoritmos de Aprendizado de máquina para prever a evasão escolar no ensino superior: um estudo no Instituto Federal de Santa Catarina.

O trabalho realizado por Primão (2022), aborda a importante questão da previsão da evasão escolar em instituições públicas de ensino superior, destacando a relevância de usar técnicas de Aprendizado de máquina para alcançar esse objetivo. O foco do estudo está no Instituto Federal de Santa Catarina (IFSC).

O principal propósito é desenvolver um método baseado em algoritmos de Aprendizado de máquina para prever a evasão escolar nessa instituição.

O estudo começa por investigar os fatores que influenciam a evasão no ensino superior, explorando a literatura existente e identificando características essenciais. Com base nisso, foi criada uma planilha de dados contendo essas características. Três algoritmos foram empregados na criação do método de previsão: Árvore de Decisão (utilizado como baseline), Rede Neural Artificial e *XGBoost.*

O estudo não apenas analisou os anos anteriores à pandemia da *Covid-19* (2017, 2018 e 2019), mas também considerou os anos de pandemia (2020 e 2021) para entender o impacto dessa situação excepcional na evasão escolar. Ambos os modelos, *XGBoost* e Rede Neural Artificial, superaram o baseline em ambas as bases de dados (antes e durante a pandemia), com *o XGBoost* demonstrando superioridade.

No conjunto de dados pré-pandemia, o algoritmo *XGBoost* atingiu um F1-Score de 97,53%, enquanto a Rede Neural Artificial alcançou 93,83%. Durante a pandemia, o *XGBoost* manteve um desempenho notável, com um *F1-Score* de 90,32%, enquanto a Rede Neural Artificial obteve um *F1-Score* de 80%.

Além disso, a análise da importância das variáveis revelou insights interessantes. Antes da pandemia, a quantidade de disciplinas concluídas foi a variável mais relevante, seguida pela forma de ingresso, média geral do discente, renda familiar per capita e campus. Durante a pandemia, a idade do discente se destacou como a variável mais importante, seguida novamente pela forma de ingresso, curso do discente, naturalidade do discente e média geral do discente.

Diante dos resultados obtidos, o algoritmo *XGBoost* foi escolhido para desenvolver um método ajustado. Esse método passou por testes de novos hiperparâmetros e a eliminação de variáveis que não demonstraram significância estatística. O método ajustado não afetou os resultados para o conjunto de dados anterior à pandemia, mas apresentou melhorias significativas para o conjunto de dados durante a pandemia.

2.6.2 Detecção de Estudantes em Risco de Evasão Escolar Usando Aprendizagem de Máquina

Neste trabalho realizado por Lopes Filho (2021), aborda um tema relevante e atual, que é a evasão escolar e seus impactos negativos. Ele menciona a abrangência do problema, afirmando que a evasão ocorre tanto na educação privada quanto na pública, e destaca os diversos perfis afetados, incluindo estudantes, instituições e o público em geral.

O texto também aponta os efeitos adversos da evasão, como o déficit educacional do aluno que evadiu, a perda financeira para o sistema educacional, o estigma social associado e os sentimentos de inadequação.

O foco do trabalho é apresentar um sistema desenvolvido para a detecção precoce de estudantes em risco de evasão, utilizando dados administrativos de alunos do ensino fundamental (anos finais) e médio de escolas públicas estaduais e superior. Ele menciona as técnicas utilizadas para análise, como *Decision Jungle, Decision Forest*, Regressão Logística *e Bayes Point Machines,* e especifica que os dados utilizados são provenientes da Secretaria de Educação do Estado de São Paulo.

Por fim, é destacado que o sistema construído é uma ferramenta de suporte à decisão que integra o cotidiano de profissionais da educação em todo o Estado, com o propósito de atender aos mais de 3,4 milhões de alunos matriculados.

2.6.3 Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica

Este artigo explora a eficiência da aplicação de técnicas de aprendizado de máquina na previsão de evasão acadêmica. Conforme Amorim (2008), O foco está na modelagem dos principais fatores que podem levar um aluno a abandonar ou trancar seu curso universitário.

O artigo apresenta as etapas principais para implementar um sistema de previsão, como a seleção de atributos, a coleta de dados e a escolha de classificadores. Três classificadores amplamente usados são avaliados quanto à acurácia, o *J48, SMO* (*Support Vector Machines) e Bayes Net*. Além disso, o artigo fornece estatísticas sobre evasão em diferentes cursos.

O artigo apresenta uma abordagem interessante ao aplicar técnicas de

aprendizado de máquina para entender a evasão acadêmica. Ele começa definindo a aprendizagem de máquina e seu papel na melhoria do desempenho em tarefas. O foco do artigo é mostrar como essas técnicas podem ser aplicadas para prever a evasão em cursos universitários.

A evasão acadêmica é introduzida como um problema significativo e complexo. O artigo discute a escassez de estudos qualitativos sobre o tema e destaca a necessidade de compreender os motivos subjacentes à evasão.

Amorim (2008) também explica de forma detalhada sobre como ocorre o aprendizado de máquina e como é útil para leitores que podem não estar familiarizados com o conceito. Os exemplos práticos, como o reconhecimento óptico de caracteres e a previsão baseada em fatores climáticos, ajudam a ilustrar como os algoritmos de aprendizado de máquina funcionam na prática.

A seleção dos classificadores *J48, SMO (Support Vector Machines) e Bayes Net* para prever a evasão é uma abordagem válida. No entanto, a análise poderia ser mais abrangente, considerando outros classificadores e discutindo porque esses três foram escolhidos.

A eficácia dos classificadores é avaliada por meio da acurácia, mostrando a proporção de classificações corretas e incorretas. É importante observar que, além da acurácia, outros indicadores de desempenho, como precisão, *recall e F1-score*, também são relevantes na avaliação de sistemas de previsão.

A seção que aborda a geração dos dados e a transformação para o formato *ARFF* é informativa, mas poderia ser mais detalhada, especialmente para leitores que não estão familiarizados com o processo técnico.

Os resultados sobre a evasão por curso são valiosos para entender a distribuição do problema em diferentes programas acadêmicos. A conclusão ressalta a importância do estudo e identifica possíveis direções futuras, incluindo a análise qualitativa dos atributos e a exploração de outros métodos de análise.

O artigo apresenta uma abordagem interessante para prever a evasão acadêmica usando técnicas de aprendizado de máquina. No entanto, algumas seções poderiam ser mais detalhadas e a análise dos classificadores poderia ser mais abrangente.

2.6.4 Motivos para evasão, vivências acadêmicas e adaptabilidade de carreira em universitários

Este estudo teve como objetivo a construção e validação da Escala de Motivos para Evasão do Ensino Superior. Ambiel (2014), informa que A construção dos itens da escala começou com uma busca por artigos em bases de dados digitais, além de entrevistas com estudantes ativos e evadidos.

A primeira versão da escala contou com 81 itens, que foram avaliados por juízes, resultando em 66 itens após revisões. Em seguida, a análise dos itens foi conduzida utilizando o modelo de *Rasch,* no qual um item foi excluído devido a ajustes fora do padrão, resultando em 65 itens restantes. Os resultados são discutidos à luz de estudos relacionados, enfatizando suas contribuições, limitações e necessidades de pesquisas futuras.

Além disso, o estudo contextualiza o problema da evasão no ensino superior no Brasil, indicando um aumento significativo no número de ingressantes nas Instituições de Ensino Superior (IES) ao longo dos anos. No entanto, a evasão continua a ser uma preocupação, afetando tanto as IES públicas quanto privadas.

Diversas causas e fatores que contribuem para a evasão são explorados, como a insatisfação com o ensino, dificuldades financeiras, falta de adaptação ao ambiente acadêmico e social, entre outros.

No campo teórico, a teoria de Tinto (1975), é destacada como uma das principais contribuições na compreensão do processo de evasão, enfatizando a integração e o engajamento do estudante na instituição. Também são discutidos outros estudos e instrumentos utilizados para avaliar a evasão, ressaltando a falta de instrumentos padronizados para avaliar os motivos específicos que levam os estudantes a abandonar seus cursos.

Nesse contexto, a Escala de Motivos para Evasão do Ensino Superior é apresentada como uma ferramenta importante para avaliar os possíveis motivos que influenciam os estudantes a considerar a possibilidade de abandonar seus cursos.

O processo de construção dos itens da escala é detalhado, incluindo a busca por artigos, entrevistas com estudantes e análise por juízes. A estrutura interna da escala é discutida em termos de sua validade e confiabilidade.

Conclui-se que, a escala preenche uma lacuna na literatura ao fornecer uma medida padronizada dos motivos potenciais que levam os estudantes a considerar a

evasão. Essa ferramenta pode ser valiosa para pesquisadores, instituições de ensino e formuladores de políticas na busca por estratégias de prevenção e acompanhamento da evasão no ensino superior.

2.6.5 Ponto de convergência entre os trabalhos

Todos os trabalhos abordam a questão da evasão escolar no contexto do ensino superior ou educação de nível similar. Cada trabalho reconhece a importância do uso de técnicas de Aprendizado de Máquina (*Machine Learning*) para prever ou entender a evasão escolar.

Eles destacam a relevância da análise de dados e a utilização de algoritmos de Aprendizado de Máquina para enfrentar esse problema.

2.6.6 Ponto de divergência entre os trabalhos

Nesta seção será apresentado alguns pontos em que os trabalhos correlatos divergem em relação a esta dissertação.

O primeiro trabalho (Primão, 2022) se concentra em prever a evasão escolar no Instituto Federal de Santa Catarina, usando algoritmos como *XGBoost* e Rede Neural Artificial. Os resultados mostram um desempenho notável desses algoritmos.

O segundo trabalho (Lopes Filho, 2021) aborda a detecção precoce de estudantes em risco de evasão, mas se concentra em escolas públicas estaduais e no ensino fundamental e médio, usando técnicas como *Decision Jungle, Decision Forest*, Regressão Logística e *Bayes Point Machines.*

O terceiro trabalho (Amorim, 2008) apresenta uma abordagem mais antiga sobre o uso de técnicas de Aprendizado de Máquina na previsão de evasão acadêmica, avaliando classificadores como *J48, SMO e Bayes Net* em cursos universitários.

O quarto trabalho (Ambiel, 2014) lida com a construção e validação da Escala de Motivos para Evasão do Ensino Superior, fornecendo uma ferramenta para avaliar os motivos que levam os estudantes a considerar a evasão.

2.6.7 Contribuições da pesquisa em relação aos trabalhos relacionados

Em comparação com o primeiro trabalho, a pesquisa proposta pode se

beneficiar da diversidade de abordagens e dados de diferentes níveis de ensino para enriquecer a análise da evasão escolar.

Em relação ao segundo trabalho, a pesquisa proposta pode trazer sua expertise em técnicas de Aprendizado de Máquina para melhorar a detecção precoce de evasão em diferentes níveis de ensino.

Comparado com o terceiro trabalho, a pesquisa pode contribuir com atualizações mais recentes de algoritmos e metodologias de Aprendizado de Máquina para previsão de evasão acadêmica.

Em relação ao quarto trabalho, a pesquisa pode complementar a abordagem qualitativa da Escala de Motivos para Evasão do Ensino Superior com a análise quantitativa de dados, permitindo uma visão mais abrangente da evasão escolar.

2.6.8 Método de referência aplicado à evasão

A aplicação da Árvore de Decisão na análise de dados e na criação de método de referência e neste caso, para a evasão universitária tem sido um tema relevante de pesquisa. A utilização desse algoritmo permite uma abordagem sistemática para entender e prever os fatores que influenciam a evasão de estudantes no ensino superior.

Ao aplicar a Árvore de Decisão nesse contexto, busca-se identificar os principais atributos que impactam a probabilidade de um aluno abandonar seus estudos. Esses atributos podem incluir características socioeconômicas, desempenho acadêmico, fatores pessoais e comportamentais, entre outros. A construção da árvore de decisão baseia-se na análise dos dados históricos e na identificação dos padrões e relações que levam à evasão.

A vantagem da utilização da Árvore de Decisão nesse cenário reside na sua capacidade de fornecer uma visão clara e interpretação direta dos fatores determinantes da evasão universitária.

A estrutura hierárquica da árvore permite a identificação de variáveis de maior importância e a compreensão das interações entre os diferentes atributos (Russell; Norvig, 2013).

Isso possibilita a criação de método de referência precisos e explicativos, que podem orientar intervenções e políticas de prevenção da evasão, com isto pretende-se neste trabalho alcançar o objetivo de propor um método de referência com

Aprendizado de máquina para previsão de evasão de alunos utilizando Árvore de Decisão.

Essa metodologia poderá contribuir significativamente para o desenvolvimento de estratégias eficazes de intervenção e tomada de decisões, visando a redução da evasão e o aumento da retenção de estudantes no ensino superior.

3 ALGORITMOS DE MACHINE LEARNIG E A ÁRVORE DE DECISÃO: A PROPOSTA DE MÉTODO DE REFERÊNCIA

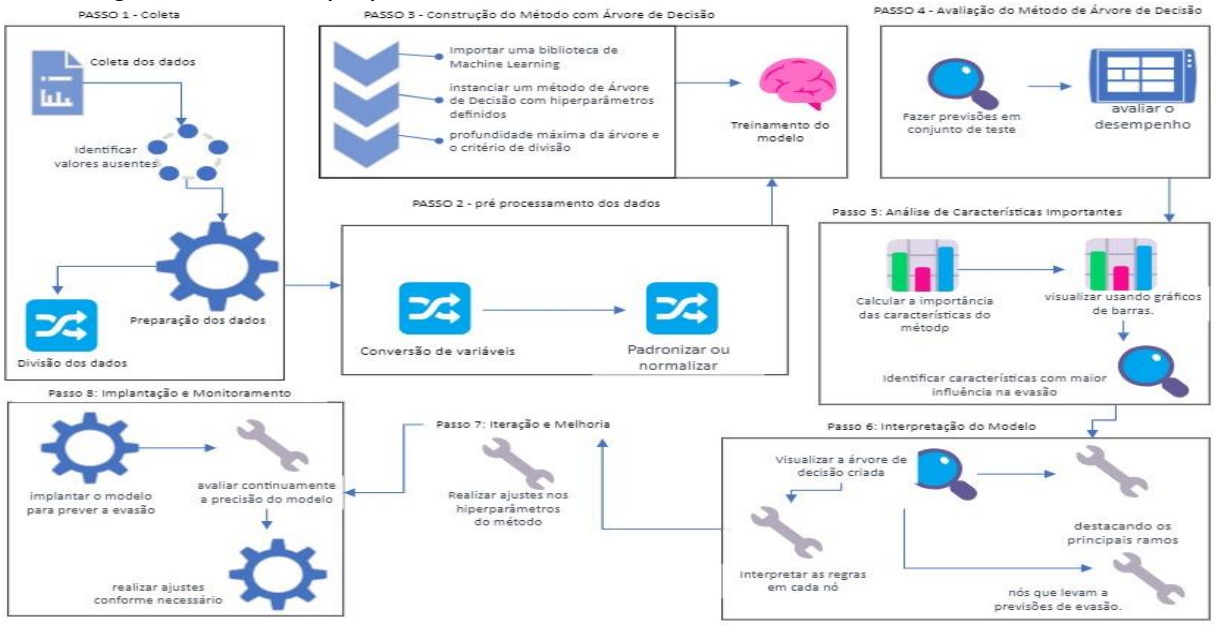
Este capítulo apresenta o método de referência para predição de alunos em risco de evasão, utilizando Aprendizado de máquina, proposto neste trabalho. A criação do método preditivo em Aprendizado de máquina da evasão dos alunos do ensino superior é realizada por meio da técnica Árvore de Decisão.

Dentre os algoritmos de Aprendizado de Máquina, a Árvore de Decisão é reconhecida como uma técnica que se baseia em uma estrutura hierárquica de decisões, semelhante a um fluxograma, em que cada nó representa uma escolha e cada ramo representa uma consequência dessa escolha. Essa característica torna o processo de 'dividir para conquistar' uma estratégia eficiente no processo decisório.

Um método de identificação antecipada de alunos com maior probabilidade de abandonar os estudos está sendo empregado. Esse método permite a intervenção e o desenvolvimento de estratégias de retenção adequadas. Em outras palavras, ele é usado para identificar alunos que têm uma maior chance de abandonar os estudos antes de concluí-los.

A utilização de técnicas de *Aprendizado de máquina*, em particular a Árvore de Decisão, oferece uma abordagem eficaz para analisar e processar grandes conjuntos de dados, considerando diversos atributos acadêmicos, sociais e pessoais dos estudantes. A Figura 9 ilustra os elementos principais do método proposto.

Figura 6 - Método proposto



Fonte: do autor (2023).

O processo de criação de um método de aprendizado de máquina envolve a preparação dos dados, a conversão dos dados para um formato adequado, a divisão dos dados em conjuntos de treinamento e teste, o treinamento e ajuste do método, a avaliação do desempenho do método, a comparação com outros métodos ou *benchmarks*, e, se necessário, ajustes adicionais para otimização.

A partir da escolha das variáveis presentes nas tabelas 2 e 3 deste estudo, elaborou-se um método de aprendizado de máquina. Este método se fundamenta na técnica de Árvore de Decisão, cuja concepção foi direcionada pela análise detalhada da Tabela 2, que descreve as variáveis selecionadas na Tabela 3.

Este processo, embasado na revisão de trabalhos científicos, visa proporcionar uma explicação mais minuciosa do escopo que será apresentado a seguir passo a passo no quadro 6:

Quadro 6 - Guia para Prevenir a Evasão Escolar no Ensino Superior com Árvore de Decisão

Passo	Nomenclatura	Classificação
1	**Coleta de Dados**	O primeiro passo envolve a coleta de dados cruciais sobre os estudantes, abrangendo variáveis como Localização da Escola, Gênero, Idade, Estado Civil, Atividade Profissional do Estudante, Educação dos Pais, Renda dos Pais, Nota média do Estudante, Pontuação no Teste Final, Pontuação no Teste Intermediário, Frequência de Atendimento às Aulas, Forma de Ingresso, Tempo desde o Ingresso, Turno de Estudo e Número de Pessoas na Família. É essencial garantir que os dados coletados estejam de alta qualidade para assegurar resultados precisos.
2	**Pré-processamento de Dados**	Na etapa seguinte, os dados coletados passam por um processo de pré-processamento. Isso inclui tratamento de valores ausentes, normalização de variáveis numéricas e codificação de variáveis categóricas. Além disso, o conjunto de dados é dividido em conjuntos de treinamento e teste para a avaliação subsequente do modelo. Esta etapa é fundamental para a confiabilidade do modelo.
3	**Construção do Método com Árvore de Decisão**	Opta-se pela árvore de decisão como método e procede-se com o treinamento do modelo utilizando o conjunto de treinamento. Ajustes nos hiperparâmetros são realizados para otimizar o desempenho. A árvore de decisão é uma escolha adequada para identificar padrões e criar regras para a detecção de evasão.
4	**Avaliação do Método**	Para avaliar a eficácia do modelo, utiliza-se o conjunto de teste. Métricas como precisão, recall, F1-score e matriz de confusão são calculadas para mensurar a capacidade do modelo em identificar casos de evasão. Ajustes são realizados conforme necessário para melhorar o desempenho.
5	**Análise das Características Importantes**	Investiga-se a importância das características no modelo, visando identificar quais fatores exercem maior influência na predição de evasão. Essa análise proporciona insights valiosos para intervenções direcionadas.
6	**Interpretação do Método**	Para compreender como o modelo toma decisões, realiza-se uma análise da estrutura da árvore de decisão. Isso permite a identificação dos caminhos que levam às previsões de evasão, viabilizando uma interpretação clara do processo de tomada de decisão.
7	**Intervenção e Aprimoramento**	Com base na análise de características e na interpretação do modelo, implementam-se estratégias de intervenção. O monitoramento contínuo e ajustes são realizados conforme necessário.
8	**Implantação e Monitoramento**	O modelo é implantado em um ambiente universitário real, estabelecendo-se um sistema de monitoramento constante para acompanhar o desempenho do modelo e a eficácia das intervenções. Mantém-se a flexibilidade para ajustar o modelo à medida que mais dados se tornem disponíveis.

Ao finalizar esses passos, o método está pronto para ser utilizado em novos dados ou tomada de decisões, e com base na descrição do quadro 9 desta seção segue uma descrição mais detalhada acerca dos dados e do processo para aplicação do método de referência.

3.1 PROCEDIMENTO PARA A ESCOLHA DAS VARIÁVEIS PARA O MÉTODO

O método de referência proposto está lastreado em uma revisão bibliográfica com o objetivo de identificar uma proposta de modelagem que pudesse contribuir para este estudo e para tal foram identificadas variáveis que foram utilizadas como preditores da evasão de alunos em diversas abordagens de maneira que, com a concatenação dos métodos eleitos, procedeu-se à catalogação de todas as variáveis relevantes nos estudos e suas relações com a predição da evasão.

Merecendo considerar que o processo de identificar variáveis relevantes e estabelecer suas relações com a predição é considerado fundamental na construção de método preditivos eficazes (Beaulac; Rosenthal, 2019).

Merece também considerar que, durante a revisão bibliográfica para a prospecção do método, foram coletadas um total de 57 variáveis em 33 estudos publicados, conforme apresentado a seguir na Tabela 2, a seguir.

Essas variáveis abrangem diferentes aspectos relacionados à evasão de alunos, como desempenho acadêmico, características demográficas, dados socioeconômicos e informações sobre a instituição de ensino. A coleta adequada de variáveis é um passo crucial para a construção de método de aprendizado de máquina eficientes e precisos (Mitchell, 1997, p. 127).

Tabela 2 - Variáveis identificadas para predição da evasão de alunos

N	Variáveis
1	Localização da Escola
2	Gênero
3	Idade
4	Estado Civil
5	O Estudante Trabalha?
6	Educação dos Pais
7	Renda dos Pais
8	Nota média do Estudante
9	Pontuação no Dever de Casa
10	Pontuação no Teste Final
11	Pontuação no Teste Intermediário
12	Frequência de Atendimento as Aulas
13	Número total de Créditos
14	Período/Semestre
15	N° de Disciplinas Aprovadas
16	N° de Disciplinas Reprovadas
17	N° de Disciplinas Canceladas
18	N° de Disciplinas Trancadas
19	Forma de Ingresso
20	Tempo em Ingresso
21	Turno
22	Carga Horária
23	Se o estudante é cotista
24	Número de pessoas na família
25	Área do Curso
26	Frequência de uso do Computador
27	Possui deficiência?
28	Número Total de Concluintes do Curso
29	Se teve experiência online anteriormente
30	Nível de Escolaridade
31	Semestre de Ingresso
32	Possui Bolsa?
33	Modalidade do Curso
34	Classificação do Curso na CAPES
35	O estudante recebe auxílio?
36	Quantidade de Aprovações e Reprovações
37	em que Tipo de instituição o estudante completou o Ensino Médio
38	Se o estudante possui residência própria
39	Se o estudante mora em Área urbana ou rural
40	Períodos Cursados
41	Coeficiente de Rendimento por Período
42	Dependências Acumuladas
43	Número Total de Matrículas do Curso
44	Número Total de Ingressantes
45	Número Total de Vagas
46	Quantidade de Computadores para alunos
47	Já fez Ensino superior antes?
48	Já fez EAD antes?
49	Local de acesso à internet
50	Tipo de conexão à internet
51	Nível de Conhecimento Informático
52	Motivo dos Estudos
53	Dificuldade
54	Se os Pais Faleceram
55	Se é superdotado
56	Número de Filhos
57	Data de Nascimento

Fonte: do autor (2023).

Importante considerar que, a análise das variáveis coletadas permitiu identificar padrões e tendências relacionados à evasão de alunos ao tempo em que merece o registro de que a identificação de padrões nos dados é um objetivo central da aprendizagem de máquina" (Mitchell, 1997, p. 56).

3.2 UMA BREVE SÍNTESE DO MÉTODO

A seção anterior apresentou uma visão abrangente do processo de preparação e aplicação de um método de Aprendizado de máquina no contexto da detecção de fatores que afetam o desempenho acadêmico dos estudantes.

Ao longo das várias etapas descritas, foi possível entender a importância da preparação adequada dos dados, da conversão para formatos compatíveis com algoritmos de aprendizado de máquina, da divisão em conjuntos de treinamento e teste, e do treinamento propriamente dito do método.

Além disso, a criação do método proposto, seu subsequente processo de avaliação e a comparação com métodos existentes foram fundamentais para estabelecer sua eficácia e posicionamento em relação a outras abordagens. Essa análise comparativa permitiu avaliar se o método desenvolvido representa uma contribuição para a área de estudo, fornecendo insights sobre o desempenho e a competitividade do método.

É importante destacar que a otimização contínua do método, com base nos resultados da comparação, é uma etapa crucial. Esse processo iterativo de ajuste visa aprimorar o desempenho e a eficácia do método, garantindo que ele esteja alinhado com os objetivos do projeto.

No contexto específico da educação, a aplicação de algoritmos de Aprendizado de máquina e árvores de decisão surge como uma abordagem para melhorar a identificação de fatores que influenciam a evasão dos estudantes. A combinação de dados e tecnologia tem o potencial de oferecer insights que podem beneficiar tanto os educadores quanto os alunos.

3.3 MÉTODO DE REFERÊNCIA APLICADO A EVASÃO

A aplicação da Árvore de Decisão na análise de dados e na criação de método de referência e neste caso, para a evasão universitária tem sido um tema relevante de pesquisa. A utilização desse algoritmo permite uma abordagem sistemática para entender e prever os fatores que influenciam a evasão de estudantes no ensino superior (Bengio, 2009).

Ao aplicar a Árvore de Decisão nesse contexto, busca-se identificar os principais atributos que impactam a probabilidade de um aluno abandonar seus estudos. Esses atributos podem incluir características socioeconômicas, desempenho acadêmico, fatores pessoais e comportamentais, entre outros. A construção da árvore de decisão baseia-se na análise dos dados históricos e na identificação dos padrões e relações que levam à evasão.

A vantagem da utilização da Árvore de Decisão nesse cenário reside na sua capacidade de fornecer uma visão clara e interpretação direta dos fatores determinantes da evasão universitária.

A estrutura hierárquica da árvore permite a identificação de variáveis de maior importância e a compreensão das interações entre os diferentes atributos (Russell; Norvig, 2013).

Isso possibilita a criação de métodos de referência precisos e explicativos, que podem orientar intervenções e políticas de prevenção da evasão, com isto pretende-se neste trabalho alcançar o objetivo de propor um método de referência com *Aprendizado de máquina* para previsão de evasão de alunos utilizando Árvore de Decisão.

3.4 RESULTADOS DA CONSULTA E CONFRONTO DAS VARIÁVEIS ELEITAS

Os trabalhos consultados reiteram que, com base nesses padrões, será possível desenvolver um método preditivo capaz de identificar alunos em risco de evasão. Nesse processo de desenvolvimento do método preditivo, diversas variáveis foram consideradas e confrontadas para obter resultados significativos.

Assim, a catalogação minuciosa de diversas variáveis, seguida por um cruzamento dessas variáveis com os trabalhos científicos já publicados permitiu a identificação de quais variáveis estavam presentes nos artigos e em outras

pesquisas relacionadas e a catalogação constou que apenas 15 variáveis foram encontradas em pelo menos em 6 trabalhos científicos simultaneamente ao tempo em que essas variáveis estão apresentadas na Tabela 3:

Tabela 3 - Variáveis selecionadas para o método

N	Variáveis
1	Localização da Escola
2	Gênero
3	Idade
4	Estado Civil
5	O Estudante Trabalha?
6	Educação dos Pais
7	Renda dos Pais
8	Nota média do Estudante
9	Pontuação no Teste Final
10	Pontuação no Teste Intermediário
11	Frequência de Atendimento as Aulas
12	Forma de Ingresso
13	Tempo em Ingresso
14	Turno
15	Número de pessoas na família

Fonte: do autor (2023).

Para compor nosso método proposto é analisado as variáveis mencionadas na Tabela 3, levando em consideração sua presença em pelo menos seis estudos de pesquisa com relevância média. Conforme destacado por Mitchell (1997, p. 87) a escolha criteriosa das características é frequentemente a etapa mais crucial no projeto de um sistema de aprendizado.

De acordo com Mitchell (1997), essa abordagem é crucial para garantir a efetividade e precisão do método e a avaliação da frequência das variáveis também está indexada ao sucesso da análise pois podemos identificar aquelas que ocorrem com maior frequência nos dados e, portanto, podem fornecer *insights* mais significativos.

Além disso, é essencial levar em conta a relevância das variáveis, ou seja, a capacidade delas de contribuir para a predição ou classificação correta. Dessa forma, ao aplicar esses princípios, podemos otimizar a construção de métodos de aprendizado de máquina, melhorando assim sua capacidade de generalização e desempenho.

3.5 CARACTERÍSTICAS DO MÉTODO PROPOSTO E VARIÁVEIS ADOTADAS

Ao analisar a performance dos estudantes, é possível considerar uma ampla gama de variáveis que podem influenciar seu desempenho e as variáveis aqui utilizadas foram elicitadas através de metodologia exploratória em estudo sistemático de obras anteriormente publicadas.

As variáveis consideradas abrangem aspectos essenciais, tais como:

- **Perfil dos Estudantes**: Localização da escola, gênero, idade, estado civil, se o estudante trabalha.
- **Contexto Familiar**: Educação dos pais, renda familiar, número de pessoas na família.
- **Desempenho Acadêmico**: Nota média do estudante, pontuações em tarefas, testes intermediários e finais, frequência de atendimento às aulas, disciplinas aprovadas, reprovadas, canceladas ou trancadas.
- **Aspectos da Instituição**: Forma de ingresso, tempo desde o ingresso, turno, carga horária, se o estudante é cotista.

Para um método ser efetivo no domínio educacional, é importante que ele seja capaz de entender conceitos e terminologias específicas do campo da educação e ter um conhecimento profundo sobre as melhores práticas de ensino e aprendizagem bem como algumas necessidades consideradas importantes, a citar as características de interpretabilidade das variáveis.

A capacidade de lidar com dados categóricos e numéricos é um dos aspectos fundamentais a serem considerados. Além disso, é importante analisar a seleção de recursos e as formas das relações, que podem ser tanto lineares quanto não lineares. A robustez a outliers, ou seja, a capacidade de lidar com dados que se distanciam radicalmente dos demais na amostra, é outra característica relevante.

Isso inclui a capacidade de detectar e tratar verdadeiros valores atípicos ou aberrantes. Tudo isso visa garantir que o modelo demande pouco pré-processamento de dados em seu uso, conforme indicado no Quadro 7.

Quadro 7 - Listas de Necessidades Importantes nas Implementações

Nomenclatura	Classificação
Interpretabilidade	A árvore de decisão é um método de aprendizado de máquina altamente interpretável, pois as decisões tomadas pelo método podem ser facilmente compreendidas e visualizadas em forma de uma estrutura hierárquica de árvore (Chui; Fung; Lok, 2020).
Lidar com dados categóricos e numéricos	A árvore de decisão pode lidar com diferentes tipos de variáveis, incluindo variáveis categóricas, como "Localização da Escola" e "Gênero", e variáveis numéricas, como "Idade", "Renda dos Pais" e "Nota média do Estudante" (Gamie; El-Seoud; Salama, 2020).
Seleção de recursos	A árvore de decisão pode ajudar na identificação das variáveis mais relevantes para a tarefa de classificação ou regressão, uma vez que as variáveis mais discriminantes tendem a estar mais próximas da raiz da árvore (Chui; Fung; Lok, 2020).
Lida com relações não lineares	A árvore de decisão é capaz de modelar relações complexas entre as variáveis de entrada e a variável de saída, mesmo quando essas relações não são lineares (Delen, 2010).
Robustez a *outliers*	A árvore de decisão é menos sensível a *outliers* em comparação com alguns outros métodos de aprendizado de máquina, pois as divisões são baseadas em limites de decisão em cada nó da árvore (Fernández; Gil; Mora, 2019).
Requer pouco pré-processamento de dados	A árvore de decisão não requer normalização ou escala dos dados de entrada, tornando o pré-processamento dos dados mais simples em comparação com alguns outros métodos (Delen, 2010).

É importante observar que a árvore de decisão pode ter a tendência de superajustar (*overfitting*) os dados de treinamento, especialmente quando a árvore se torna muito profunda ou complexa. Portanto, o ajuste adequado de hiperparâmetros e técnicas de regularização, como a poda da árvore, podem ser necessários para controlar o *overfitting* e melhorar o desempenho geral dos métodos.

Também merece considerar que, além das características específicas do método considerando o uso da técnica de Árvore de Decisão mencionadas anteriormente, existem 6 outras características relevantes para que precisa ser

mencionada devido a natureza do trabalho conforme quadro 8:

Quadro 8 - Características Relevantes do Método Considerado

Nomenclatura	Caracterísiticas Relevantes
Facilidade de interpretação dos resultados	A Árvore de Decisão permite que os resultados sejam facilmente interpretados por diferentes stakeholders, como pesquisadores, professores e administradores universitários, facilitando a compreensão dos fatores que influenciam o desempenho dos estudantes (Chui; Fung; Lok, 2020).
Identificação de padrões e *insights*	A Árvore de Decisão pode revelar padrões e relações entre as variáveis que podem não ser facilmente identificados por outros métodos analíticos. Isso pode ajudar a descobrir *insights* valiosos sobre os fatores que afetam o sucesso acadêmico dos estudantes (Fernández; Gil; Mora, 2019).
Identificação de variáveis-chave	A Árvore de Decisão pode destacar as variáveis mais importantes que têm um impacto significativo no desempenho dos estudantes. Isso pode auxiliar na alocação eficiente de recursos e na identificação de áreas em que é necessário intervir para melhorar os resultados acadêmicos (Chui; Fung; Lok, 2020).
Segmentação de estudantes	Com base nas divisões feitas pela Árvore de Decisão, é possível segmentar os estudantes em grupos ou subgrupos com características semelhantes. Isso permite uma análise mais aprofundada de cada grupo e a identificação de estratégias específicas para melhorar o desempenho de cada segmento (Delen, 2010).
Monitoramento contínuo	A árvore de decisão pode ser atualizada à medida que novos dados são disponibilizados, permitindo o monitoramento contínuo do desempenho dos estudantes ao longo do tempo. Isso pode ajudar a identificar tendências e mudanças que ocorrem ao longo do período analisado (Gamie; El-Seoud; Salama, 2020).
Apoio à tomada de decisão	Os resultados da análise com árvores de decisão podem fornecer uma base sólida para a tomada de decisões na área educacional. Isso pode incluir a implementação de programas de apoio específicos, a adaptação de estratégias de ensino ou o desenvolvimento de políticas voltadas para melhorar os resultados acadêmicos dos estudantes (Géron, 2019).

Neste estudo, apresenta-se um método de referência que utiliza a técnica de

Aprendizado de Máquina, especificamente a Árvore de Decisão, com o objetivo de antecipar a identificação dos fatores que contribuem para a evasão de alunos em instituições de ensino.

Essa abordagem fornece uma base sólida para pesquisas futuras e viabiliza a extração de informações de bases de dados históricas. Isso, por sua vez, permite que as equipes de gestão implementem estratégias de prevenção de evasão com antecedência.

A Árvore de Decisão é notável por sua capacidade de apresentar resultados de forma acessível, identificar padrões relevantes, selecionar variáveis-chave, segmentar estudantes em grupos similares e facilitar a tomada de decisões informadas.

3.6 TÉCNICAS DE APLICAÇÃO DAS ÁRVORES DE DECISÃO

Nesta seção, são exploradas as técnicas de aplicação das árvores de decisão, ferramentas amplamente utilizadas devido à sua eficácia e simplicidade. Essas árvores oferecem uma estrutura hierárquica que representa decisões e testes em atributos, permitindo a tomada de decisões intuitiva. Além disso, são fundamentais para explicar o raciocínio por trás das decisões de aprendizado de máquina.

A integração das árvores de decisão em sistemas complexos possibilita justificar ações de inteligência artificial, enquanto o uso de ensemble de árvores, como *Random Forest* ou *Gradient Boosting*, oferece robustez e precisão, permitindo a interpretação da importância dos atributos nas decisões. As árvores de decisão desempenham um papel crucial na compreensão e transparência das ações da inteligência artificial.

3.7 OS NÓS INTERNOS

Importante verificar que cada nó interno da árvore de decisão são estruturas hierárquicas que representam um conjunto de decisões a serem tomadas e representam um teste em um atributo específico, enquanto as arestas correspondem aos possíveis resultados desse teste.

Os nós da folha da árvore representam as classes ou as saídas desejadas de

forma que essa estrutura oferece uma abordagem intuitiva e transparente para tomar decisões com base em dados.

As decisões que são inicialmente tomadas por máquinas precisam ser recompensadas a humanidade (Russell; Norvig, 2013), assim nesse contexto, o uso de árvores de decisão pode fornecer um mecanismo para explicar o pensamento por trás das decisões tomadas por um método de aprendizado de máquina.

3.8 AS TÉCNICAS UTILIZADAS E AS CAMADAS INTERPRETÁVEIS

Sobre as técnicas utilizadas no processo decisório merece o registro de que ao utilizar árvores de decisão como uma camada interpretável em sistemas mais complexos, é possível fornecer justificativas compreensíveis para as ações tomadas pela inteligência artificial (Gomes, 2021).

A exemplo tem-se uma técnica comum que é o uso de ensemble de árvores de decisão, como o *Random Forest* ou *Gradient Boosting*, que, em si proporciona um ensaio matemático de processo de criação de muitas árvores de decisão, de maneira aleatória, formando o que podemos enxergar como uma floresta, onde cada árvore é utilizada na escolha do resultado final, em uma espécie ou categoria de votação (Mitchell, 1997).

Essas combinações variam de acordo com um método mais robusto e preciso, efetivamente a generalização e a capacidade de lidar com dados complexos de maneira que, além disso, permitem avaliar a importância dos atributos no processo de tomada de decisão. Isso permite que os usuários compreendam o pensamento por trás das decisões tomadas pelo método de aprendizado de máquina (Russell; Norvig, 2013).

3.9 A ÁRVORE DE DECISÃO COMO COMO UMA CAMADA INTERPRETÁVEL

Ao utilizar Árvore de Decisão como uma camada interpretável, é possível fornecer justificativas compreensíveis para as ações tomadas pela inteligência artificial (Gomes, 2021) e ao analisar a perspectiva conforme preconizado por Russell e Norvig (2013), pode-se destacar as seguintes vantagens da aplicação de Árvore de Decisão no contexto do trabalho:

Quadro 9 - Vantagens da aplicação de árvores de decisão no contexto do aprendizado de máquina

Nomenclatura	Classificação
Interpretabilidade	As árvores de decisão fornecem uma estrutura clara e intuitiva para explicar como as decisões são tomadas com base nos dados de entrada. Isso é especialmente importante quando a transparência e a justificativa das decisões são essenciais, como em áreas críticas, como saúde e direito (Fernández; Gil; Mora, 2019).
Facilidade de Implementação	As árvores de decisão são relativamente fáceis de implementar e entender. Além disso, existem bibliotecas e ferramentas disponíveis que simplificam sua construção e análise (Delen, 2010).
Eficiência Computacional	Eficiência Computacional: Comparadas a métodos mais complexos, as árvores de decisão tendem a ter um desempenho computacionalmente mais eficiente, tornando-as ideais para lidar com grandes volumes de dados em tempo real (Gamie; El-Seoud; Salama, 2020).
Combinação com Outras Técnicas	As árvores de decisão podem ser usadas como parte de um conjunto de métodos, combinando-as com outras técnicas de aprendizado de máquina, como redes neurais, para obter melhorias de desempenho e generalização (Chui; Fung; Lok, 2020).

Também é importante considerar que a interpretabilidade oferecida por essas estruturas permite entender e justificar as decisões tomadas pelos métodos, promovendo maior confiança e transparência de forma que ao combinar a eficiência computacional e a facilidade de implementação das árvores de decisão com outras técnicas avançadas, é possível obter soluções mais robustas e precisas.

De acordo com Bengio (2013), a classificação utilizando árvores de decisão é uma técnica amplamente utilizada para categorizar estudantes com base em variáveis como desempenho acadêmico, nível de satisfação e probabilidade de evasão.

Essa abordagem permite uma compreensão mais aprofundada dos fatores que influenciam o sucesso ou fracasso dos estudantes, e, além disso, Bengio (2013) destaca que as árvores de decisão também podem ser aplicadas em análises de regressão, permitindo prever variáveis contínuas, como a nota final do estudante ou o tempo necessário para concluir um determinado curso de forma que essa análise de regressão fornece *insights* sobre os fatores que têm um impacto direto no desempenho acadêmico.

3.10 A ANÁLISE DE SEGMENTAÇÃO

De acordo com Mitchell (1997), a análise de segmentação é outra técnica aplicável utilizando árvores de decisão pois ao segmentar os estudantes em grupos ou subgrupos com características semelhantes, é possível analisar diferentes perfis de estudantes e identificar estratégias específicas para cada grupo.

Conforme Mitchell (1997), menciona que é possível identificar um grupo de estudantes com alto potencial acadêmico, mas baixo envolvimento em atividades extracurriculares. Essa identificação pode indicar a necessidade de incentivos adicionais para aumentar a participação desses estudantes em atividades fora da sala de aula.

3.11 A SELEÇÃO DE RECURSOS E SUA IMPORTÂNCIA PARA O CASO DA ANÁLISE DE DADOS UNIVERSITÁRIOS

A seleção de recursos é uma outra aplicação importante das árvores de decisão no contexto da análise de dados universitários. Conforme preconiza Bengio (2013), essa técnica auxilia na identificação das variáveis mais relevantes e importantes para o desempenho dos estudantes.

Dessa forma, é possível priorizar esforços e recursos em áreas específicas que têm um impacto significativo nos resultados acadêmicos.

A análise de sensibilidade, mencionada por Mitchell (1997), é uma técnica que permite avaliar o impacto das mudanças nas variáveis de entrada nos resultados obtidos pelas árvores de decisão pois com isso, é possível compreender como alterações em características específicas dos estudantes podem afetar suas chances de sucesso acadêmico.

Portanto, a utilização das técnicas de classificação, regressão, análise de segmentação, seleção de recursos e análise de sensibilidade com árvores de decisão proporciona uma abordagem abrangente e valiosa na análise de dados universitários, permitindo uma compreensão mais detalhada dos fatores que influenciam o desempenho acadêmico dos estudantes (Bengio, 2013; Mitchell, 1997).

3.12 APLICAÇÃO NO MÉTODO PROPOSTO

A evasão de alunos nas instituições de ensino superior é uma preocupação constante para os gestores acadêmicos. A perda de estudantes durante o curso pode afetar a reputação da instituição, além de representar uma perda financeira significativa. Para lidar com esse desafio, propomos a aplicação de um método de referência com Aprendizado de máquina para previsão de evasão de alunos.

3.12.1 Validação do Método

O objetivo de coordenadores (profissionais) responderem um questionário sobre suas percepções sobre evasão é coletar dados qualitativos e subjetivos que ajudem a compreender a visão desses profissionais em relação aos fatores que contribuem para a evasão escolar.

Esse tipo de pesquisa qualitativa pode fornecer informações valiosas sobre as percepções, experiências e insights dos profissionais que estão envolvidos no sistema educacional.

A relação deste questionário com a técnica de método de referência empregada está relacionada à coleta de informações para identificar os fatores que levam à evasão de alunos nas instituições de ensino.

Ao coletar as respostas dos coordenadores por meio de perguntas fechadas com escalas de 1 a 5, os pesquisadores podem quantificar as percepções desses profissionais em relação à influência de diferentes fatores na evasão, que variava o seu valor de 1 a 5, sendo descrito da seguinte forma:

1 - Nenhuma influência na evasão,

2 - Pouca influência na evasão,

3 - Influência média na evasão,

4 - Influência moderada na evasão,

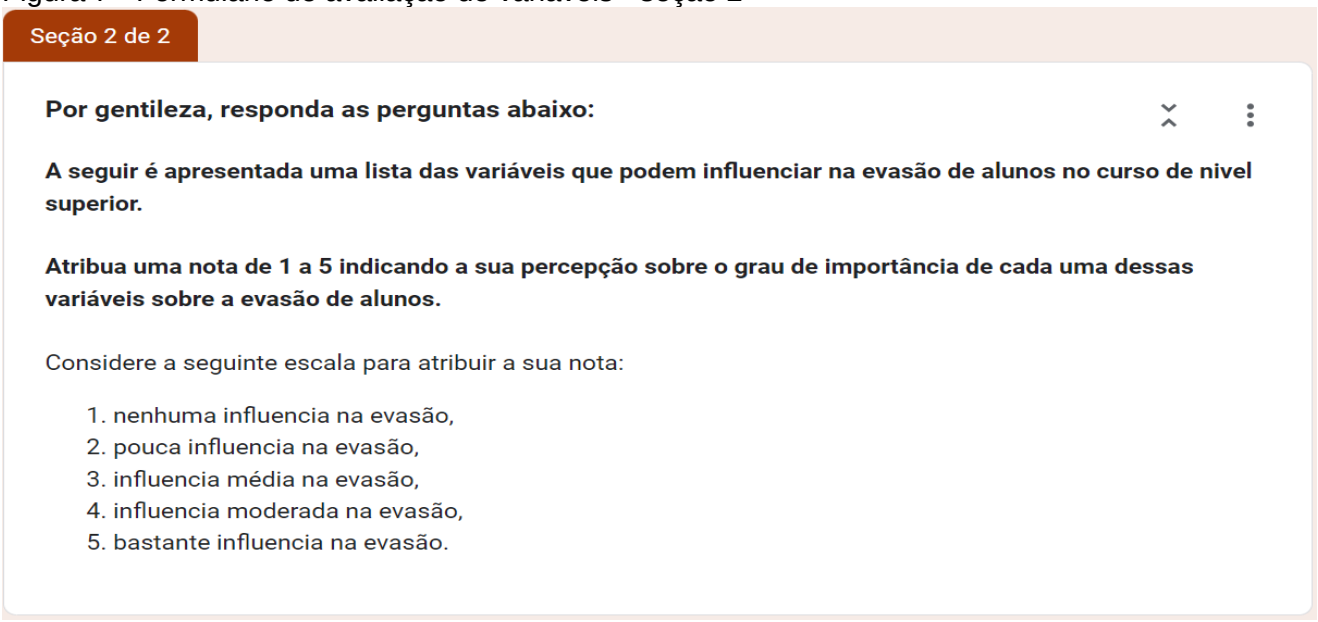
5 - Bastante influencia na evasão.

Essas respostas podem ser usadas como dados de entrada no método de referência, como a Árvore de Decisão mencionada anteriormente, para ajudar a criar um modelo preditivo ou analítico que identifique padrões ou variáveis-chave que influenciam a evasão.

Em outras palavras, as respostas dos coordenadores ajudam a alimentar o modelo analítico ou preditivo, contribuindo para uma abordagem mais abrangente e informada na prevenção da evasão escolar.

Conforme a Figura 10, que também pode ser encontrada com maiores detalhes em nosso apêndice A:

Figura 7 - Formulário de avaliação de variáveis - seção 2



Fonte: Formulário do google.

O questionário de aplicação foi dividido em 2 seções, onde na seção 1 Foi perguntado aos participantes seu email, se trabalhava também em instituição pública e qual área do curso que atuava.

Na seção 2, foi solicitada a lista das variáveis que podem influenciar a evasão de alunos no curso de nível superior, conforme apresentado na Tabela 3 - Variáveis selecionadas para o método nesta seção.

Em relação a essa questão, indagamos aos participantes se trabalham em uma instituição pública ou privada. Os resultados revelaram que a maioria, representando 90,9% dos entrevistados, trabalha exclusivamente em instituições privadas. Por outro lado, 9,1% dos participantes possuem uma visão mais abrangente do processo, uma vez que pertencem a instituições públicas.

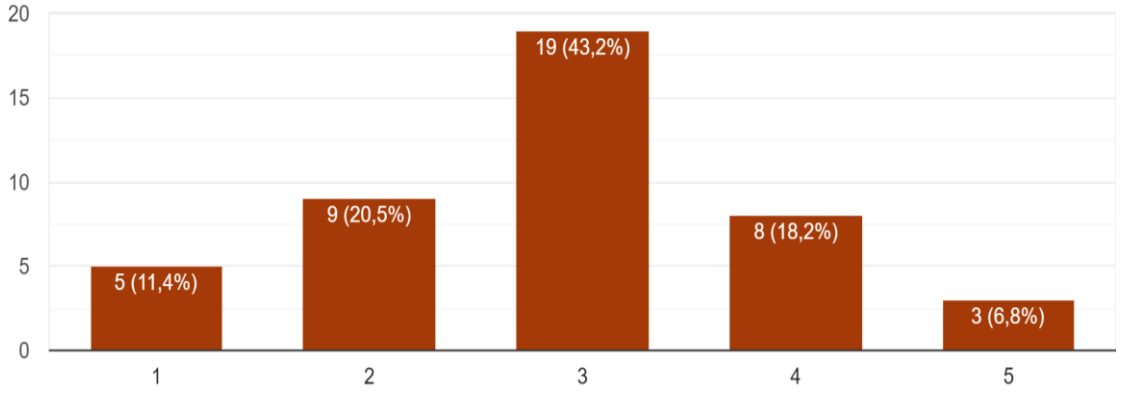
Ao perguntarmos aos profissionais em quais áreas atuavam foi recebido 59,1% pertencem a área de exatas, 29,5% estão na área de humanas e 2,3% na área de saúde, restando 9,1% para área de biologia e contábeis.

3.12.2 Sobre a localização da escola e gênero

Foi perguntando aos participantes se a localização da escola é fator relevante para que ocorra a evasão universitária, 43,2% dos colaboradores responderam que esta variável tem influência média na evasão, conforme gráfico a seguir.

Conforme figura 11, os demais resultados foram: 1 - nenhuma influência na evasão (11,4%), 2 - pouca influência na evasão (20,5%), 3 - influência média na evasão (43,2%), 4 - influência moderada na evasão (18,2%), 5 - bastante influência na evasão (6,8%).

Figura 11 - Localização da Escola

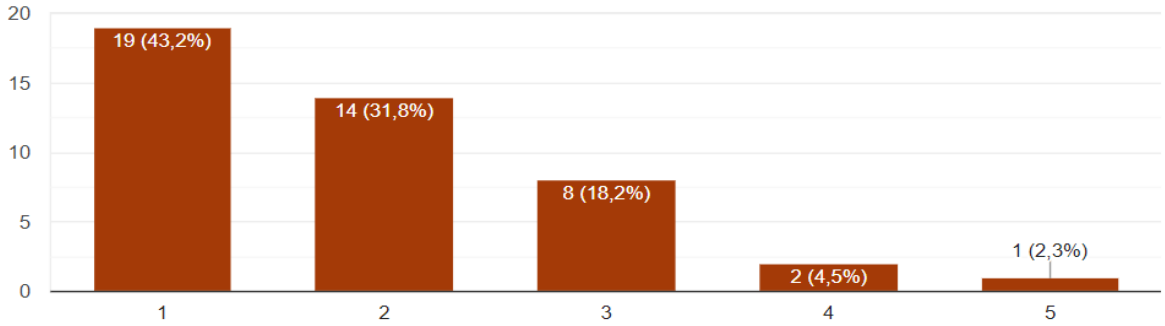


Fonte: do autor (2023).

Em seguida, ao serem questionados sobre o efeito do gênero na evasão universitária, quase metade dos participantes expressaram opiniões em que o gênero não tem influência significativa ou possui uma influência mínima nessa questão conforme dados demonstrados na figura 12.

Os dados coletados revelaram as seguintes porcentagens em relação a essa variável: 1 - nenhuma influência na evasão - 43,2%, 2 - pouca influência na evasão - 31,8%, 3 - influência média na evasão - 18,2%, 4 - influência moderada na evasão - 4,5%, 5 - bastante influência na evasão - 2,3%.

Figura 11 - Gênero



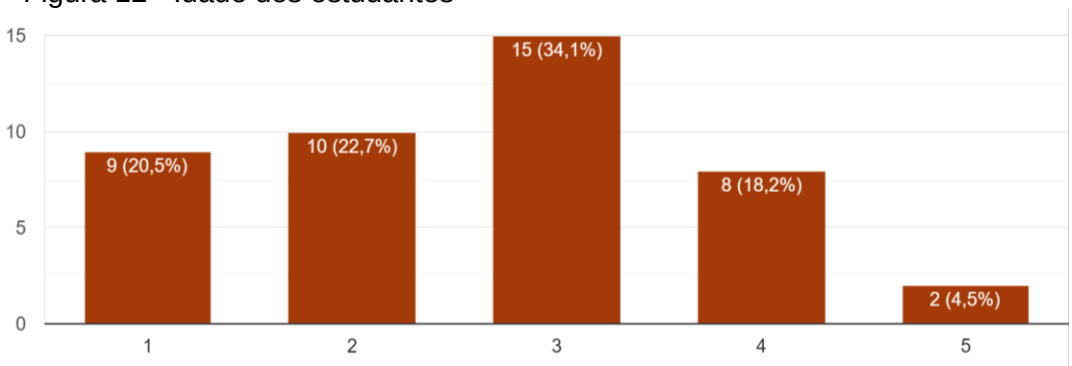
Fonte: do autor (2023).

3.13.3 Sobre a idade dos estudantes

Dando continuidade no questionário, foi perguntado sobre as demais variáveis e seus impactos e relevâncias no processo de evasão universitária e foi recebido as seguintes conforme figura 13:

A idade dos estudantes mostrou um valor de 34,1% de destaque médio no impacto da evasão universitários e apenas 4,5% entre os entrevistados informam que esta variável tem impacto relevante na evasão universitária.

Figura 12 - Idade dos estudantes



Fonte: do autor (2023).

A análise dos dados revelou que a idade dos estudantes desempenha um papel significativo no índice de evasão universitária, apresentando um valor de destaque médio de 34,1%. No entanto, surpreendentemente, apenas 4,5% dos entrevistados afirmaram que essa variável possui um impacto relevante na evasão universitária.

Esses resultados levantam questionamentos sobre a percepção dos entrevistados em relação à influência da idade na decisão dos estudantes de

abandonarem seus cursos. Enquanto a análise estatística aponta para uma associação considerável entre a idade e a evasão, é interessante notar que essa relação não é amplamente reconhecida pelos participantes do estudo.

A idade é um fator que pode influenciar diversos aspectos da vida acadêmica, como maturidade, responsabilidades pessoais e até mesmo motivação para persistir nos estudos.

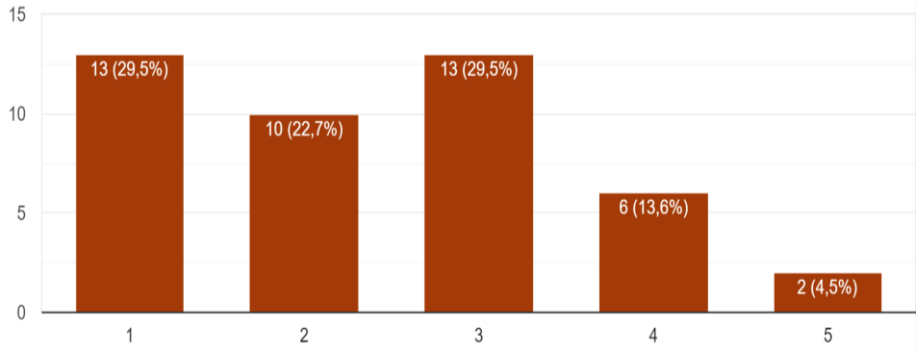
Embora os números mostrem que a idade tem um impacto notável na evasão universitária, a discrepância entre os dados estatísticos e a percepção dos entrevistados sugere a necessidade de uma reflexão mais profunda sobre as razões subjacentes a essa disparidade, o que sustenta mais ainda o presente trabalho no desenvolvimento de um método de referência.

3.12.4 Sobre o estado civil

Outra análise importante em nosso método de referência é a variável estado civil, pois ela demonstra um impacto médio na relevância da evasão universitária com 29,5% e tendo o mesmo valor para nenhuma influência para a evasão de acordo a análise dos colaboradores.

Esses dados sugerem que o estado civil pode ser um elemento a ser considerado ao estudar a evasão universitária, pois não é apresentado um valor padrão nas respostas recebidas.

Figura 13 - Estado Civil



Fonte: do autor (2023).

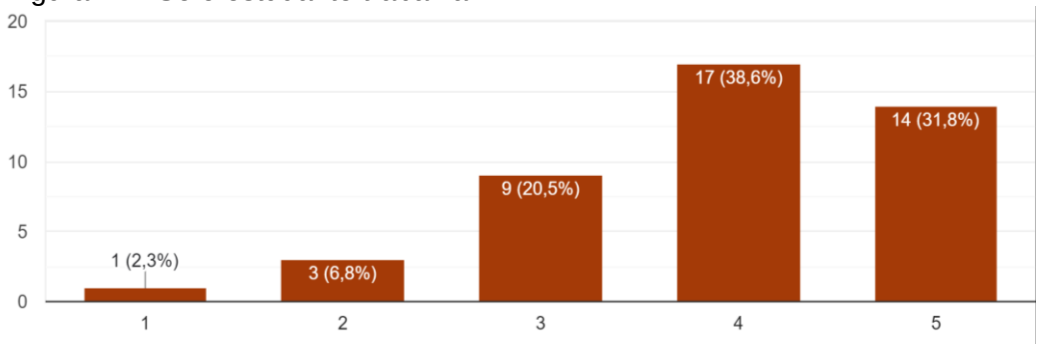
Ao analisar a variável - se o estudante trabalha, foi constatado que tem impacto relevante para o processo decisório dos alunos em relação a decidir evadir dos cursos universitários. O gráfico da variável de referência relacionada à influência

do trabalho na evasão universitária apresenta um padrão interessante.

3.12.5 Sobre a atuação em trabalho

Com base nas percentagens fornecidas expostas no gráfico 5, pode-se observar que a maioria dos coordenadores (38,6%) acreditam que o trabalho exerce uma influência moderada na evasão. Isso indica que esses estudantes percebem que conciliar as demandas do trabalho e dos estudos pode ser desafiador, e em muitos casos ocorrem a evasão.

Figura 14 - Se o estudante trabalha



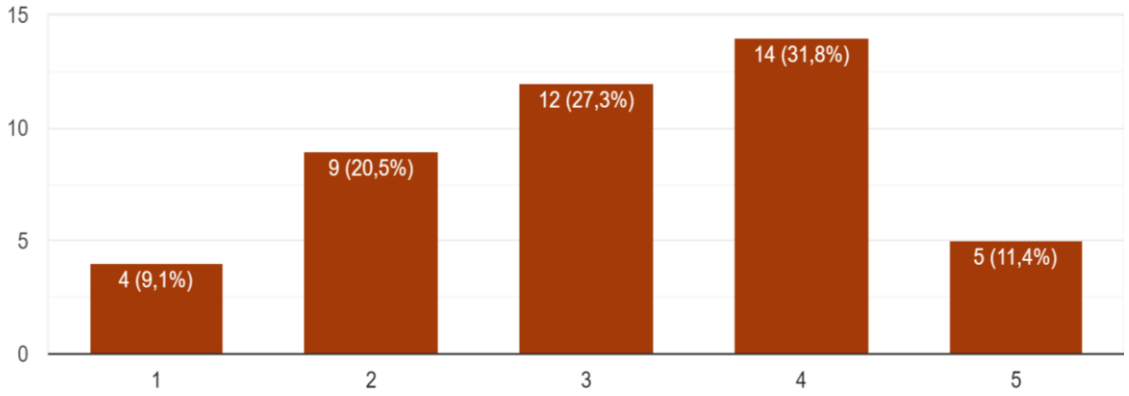
Fonte: do autor (2023).

3.12.6 Referente a educação dos pais

Após isso, procedeu-se à obtenção do valor de votação para a variável "educação dos pais", a fim de analisar sua relevância e impacto no processo de tomada de decisão, sob a perspectiva do coordenador de cursos.

Há uma significativa oportunidade a ser explorada nessa variável em nosso trabalho, pois, com base nos valores atribuídos pelos entrevistados, eles indicam que não há consenso quanto à aplicação dessa variável, de forma que ela possa influenciar no processo de evasão dos estudantes do ensino superior.

Figura 15 - Educação dos Pais



Fonte: do autor (2023).

Observou-se que os valores recebidos nesta variável conforme figura 16 foram: 1 - nenhuma influência na evasão - 9,1%, 2 - pouca influência na evasão - 20,5%, 3 - influência média na evasão 27,3%, 4 - influência moderada na evasão - 31,8%, 5 - bastante influencia na evasão - 11,4%.

Através da catalogação destes valores percebe-se que a influência está entre média e moderada no processo decisões sobre o olhar dos coordenadores de cursos.

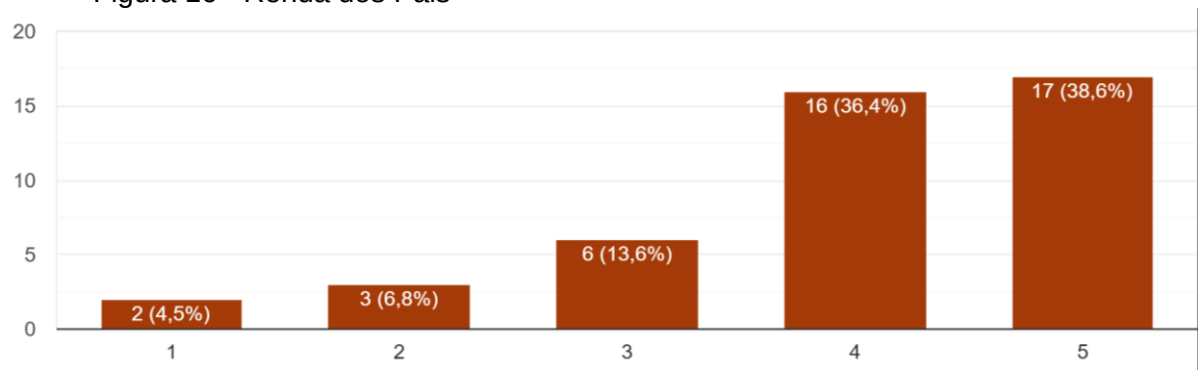
Outra variável importante a ser considerada nesta análise é a renda dos pais, uma vez que, de acordo com as informações fornecidas pelos coordenadores entrevistados, muitos jovens que estão cursando o ensino superior enfrentam dificuldades para arcar com as mensalidades em instituições privadas.

Eles dependem da renda de seus pais ou responsáveis financeiros, e caso ocorra uma alteração nessa situação, a probabilidade de os alunos trancarem o curso ou até mesmo evadirem é significativa.

3.12.7 Sobre a renda dos pais

Nesse contexto, os valores que foram atribuídos a esta variável conforme gráfico 7, foram: 1 - nenhuma influência na evasão (4,5%), 2 - pouca influência na evasão (6,8%), 3 - influência média na evasão (13,6%), 4 - influência moderada na evasão (36,4%), 5 - bastante influência na evasão (38,6%).

Figura 16 - Renda dos Pais



Fonte: do autor (2023).

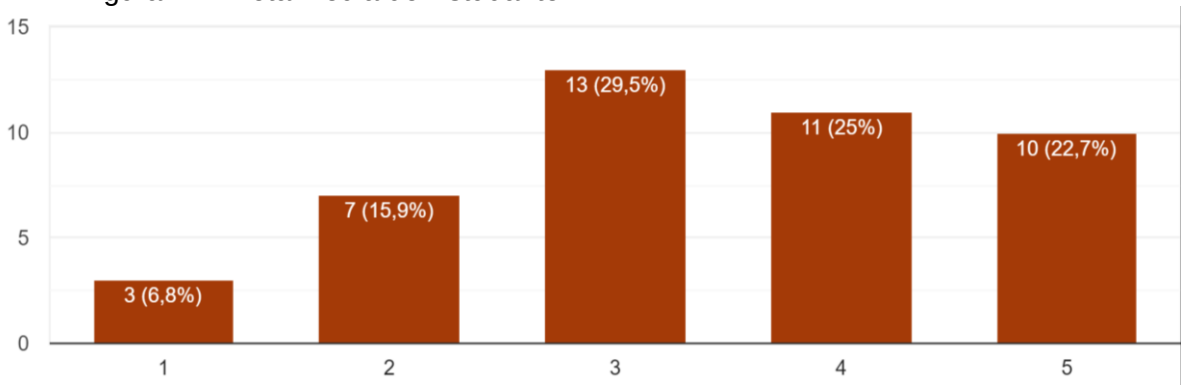
3.12.8 Sobre a média das notas do estudante

A nota média do estudante é uma variável de grande importância no processo decisório de evasão de um curso de nível superior. A média das notas obtidas ao longo do curso reflete o desempenho acadêmico do estudante e pode influenciar diretamente sua motivação, persistência e chances de sucesso.

Conforme figura 18, esta variável, detém uma moderada influência e pode causar a evasão do aluno de suas classes de nível superior. Contudo, é importante ressaltar que a nota média não deve ser o único fator a ser considerado no processo de tomada de decisão de evasão. Outros aspectos, como interesse pelo curso, satisfação com a área de estudo, adequação ao ambiente acadêmico e perspectivas de carreira, também devem ser levados em conta.

Neste sentido os valores que foram atribuídos a esta variável contidos na figura 18, foram: 1 - nenhuma influência na evasão (6,8%), 2 - pouca influência na evasão (15,9%), 3 - influência média na evasão (29,5%), 4 - influência moderada na evasão (25%), 5 - bastante influência na evasão (22,7%).

Figura 17 - Nota Média do Estudante



Fonte: do autor (2023).

Uma nota média baixa pode indicar dificuldades de aprendizado, falta de interesse ou problemas pessoais que estejam afetando o desempenho do estudante. Essa situação pode levar à desmotivação, descrença nas próprias capacidades e, consequentemente, à evasão do curso.

Por outro lado, uma nota média alta demonstra um bom rendimento acadêmico, dedicação e comprometimento do estudante com os estudos. Isso pode fortalecer a confiança e a motivação para continuar avançando no curso, pois o estudante enxerga resultados positivos do seu esforço.

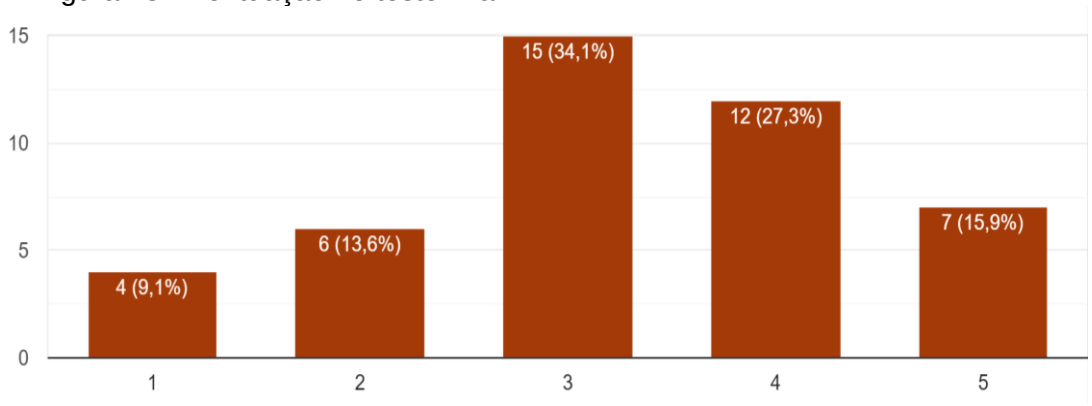
3.12.9 Sobre a pontuação no teste final

A pontuação no teste final é um aspecto crucial para o processo decisório de um estudante evadir do curso de nível superior. Ela representa o desempenho do aluno ao longo do período letivo e é geralmente utilizada como um indicador da compreensão e assimilação dos conteúdos abordados durante as disciplinas.

A maioria dos coordenadores entrevistados conforme figura 19, apresentaram uma opinião que esta variável tem impacto médio a moderado no processo decisório de evasão universitária.

DE acordo com a figura gráfico 19 que expõe as notas aplicadas pelos coordenadores que foram: 1 - nenhuma influência na evasão (6,8%), 2 - pouca influência na evasão (13,6%), 3 - influência média na evasão (34,1%), 4 - influência moderada na evasão (27,3%), 5 - bastante influência na evasão (15,9%).

Figura 18 - Pontuação no teste final



Fonte: do autor (2023).

A pontuação obtida no teste final reflete não apenas o conhecimento adquirido, mas também a capacidade de aplicar esse conhecimento de forma eficiente e eficaz. É um reflexo do esforço dedicado aos estudos, à participação ativa nas aulas e ao desenvolvimento das habilidades necessárias para o bom desempenho acadêmico.

A relevância da pontuação no teste final para o processo decisório de evasão se baseia no pressuposto de que um baixo desempenho acadêmico pode indicar problemas subjacentes que podem levar um estudante a desistir do curso.

Dificuldades de compreensão dos conteúdos, falta de motivação, problemas pessoais ou financeiros são apenas alguns exemplos de situações que podem levar um aluno a considerar a evasão.

3.12.10 Sobre a pontuação no teste intermediário

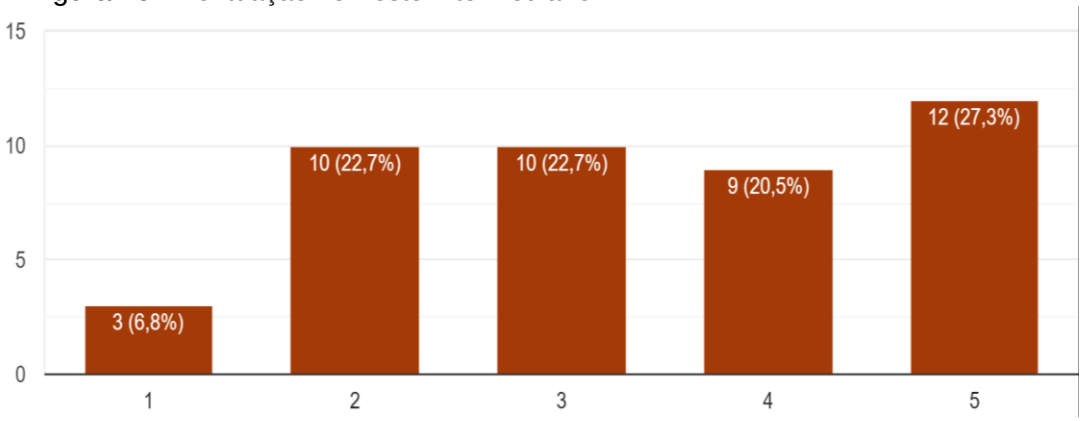
Em seguida, foi analisada a variável pontuação no teste intermediário exposta no gráfico 10 apresentado total disparidade em sua análise realizada pelos coordenadores de curso, pois não se observou nenhum valor significativo e discrepante registrado nesta pesquisa.

A pontuação obtida nesse teste pode ser uma ferramenta valiosa para que o estudante avalie seu próprio progresso e os coordenadores possam identificar possíveis dificuldades enfrentadas pelos alunos.

Ao analisar sua pontuação, o aluno pode perceber se está acompanhando o ritmo da disciplina e se possui o conhecimento necessário para continuar avançando no curso. Caso a pontuação seja baixa, isso pode indicar a necessidade de reforçar

determinados conteúdos, buscar ajuda extra ou adotar estratégias de estudo mais eficientes.

Figura 29 - Pontuação no Teste Intermediário



Fonte: do autor (2023).

Conforme figura 20 que nos mostra os valores recebidos dos coordenadores que foram: 1 - nenhuma influência na evasão (6,8%), 2 - pouca influência na evasão (22,7%), 3 - influência média na evasão (22,7%), 4 - influência moderada na evasão (20,5%), 5 - bastante influência na evasão (27,3%).

O teste intermediário é uma avaliação que ocorre durante o percurso acadêmico, geralmente no meio do semestre ou ano letivo, e tem como objetivo verificar o conhecimento e o desempenho dos alunos em relação aos conteúdos abordados até aquele momento.

3.12.11 Sobre frequência no atendimento às aulas

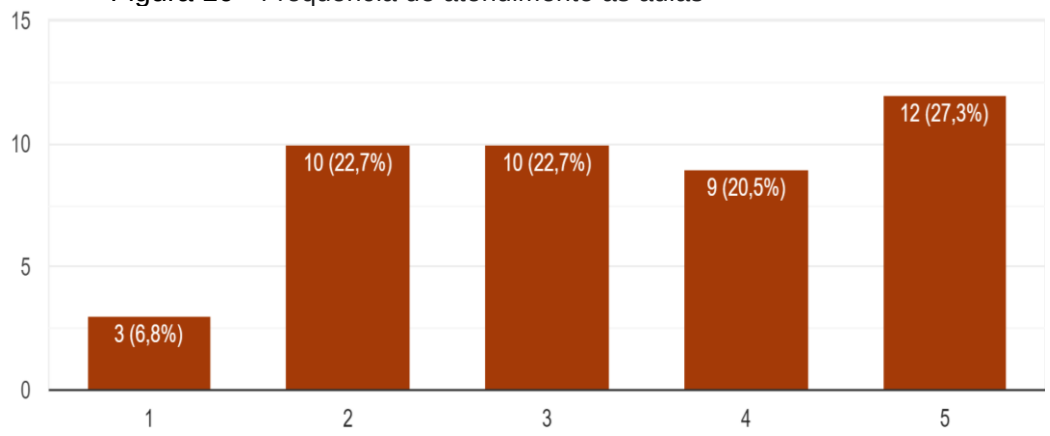
A frequência no atendimento às aulas é um fator relevante no processo de evasão dos alunos de nível superior. Quando analisamos essa variável, percebemos que a falta de um valor padronizado entre os coordenadores de curso dificulta uma análise assertiva e com resultados satisfatórios. Portanto, é necessário aplicar um método de referência que permita uma avaliação mais precisa.

A frequência no atendimento às aulas reflete o comprometimento do estudante com seu curso e suas responsabilidades acadêmicas. Alunos que frequentam regularmente as aulas demonstram maior engajamento e interesse na obtenção de conhecimento.

Essa presença assídua contribui para um melhor aproveitamento das

disciplinas, permitindo a participação ativa em discussões e atividades práticas.

Figura 10 - Frequência de atendimento as aulas



Fonte: do autor (2023).

A ausência frequente às aulas também pode levar a um distanciamento do estudante em relação aos colegas e professores, resultando em uma menor integração social no ambiente universitário.

A interação com os colegas de classe e a participação nas atividades extracurriculares são elementos importantes para a formação acadêmica e pessoal do aluno. A falta de envolvimento nessas atividades pode levar a um sentimento de isolamento, influenciando negativamente o seu desejo de permanecer no curso.

Porém, ao analisar os valores recebidos pelos entrevistados nos mostra que esta variável não detém um valor significativo e que mostre relevância no processo decisório de evasão entre os universitários.

Observa-se o figura 21 que nos mostra os valores registrados pelos coordenadores de curso que foram: 1 - nenhuma influência na evasão (6,8%), 2 - pouca influência na evasão (22,7%), 3 - influência média na evasão (22,7%), 4 - influência moderada na evasão (20,5%), 5 - bastante influência na evasão (27,3%).

O teste intermediário é uma avaliação que ocorre durante o percurso acadêmico, geralmente no meio do semestre ou ano letivo, e tem como objetivo verificar o conhecimento e o desempenho dos alunos em relação aos conteúdos abordados até aquele momento.

3.12.12 Sobre a forma de ingresso ao curso

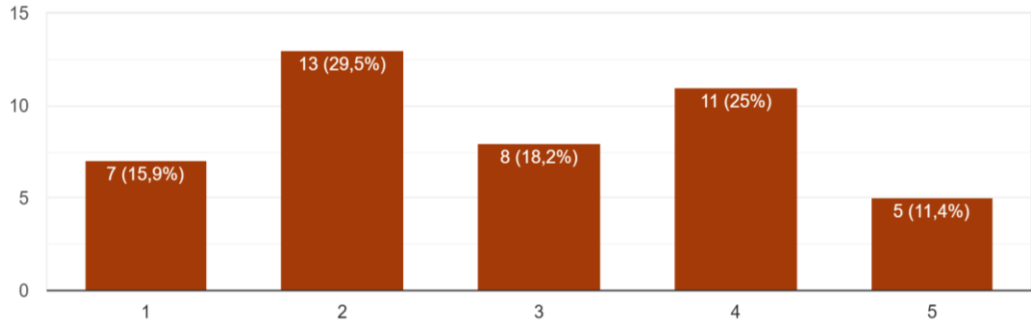
Outra variável que foi abordada em nosso trabalho foi formar de ingresso

universitário, existem diferentes formas de ingresso em uma instituição de ensino superior, como vestibular, SISU (Sistema de Seleção Unificada), ENEM (Exame Nacional do Ensino Médio), cotas e processos seletivos específicos, como as provas de habilidades específicas para cursos de artes, por exemplo.

Cada uma dessas formas de ingresso tem suas peculiaridades e critérios de seleção, o que pode afetar diretamente a motivação e o sentimento de pertencimento do estudante.

A variável "formas de ingresso" é um aspecto essencial quando se trata do processo decisório de um estudante em evadir de um curso de nível superior. A forma como um aluno ingressa em uma instituição de ensino superior pode influenciar sua experiência acadêmica e seu engajamento com o curso.

Figura 11 - Formas de ingresso



Fonte: do autor (2023).

De acordo com a figura 22 que nos mostra os valores recebidos dos coordenadores que foram: 1 - nenhuma influência na evasão (15,9%), 2 - pouca influência na evasão (29,5%), 3 - influência média na evasão (18,2%), 4 - influência moderada na evasão (25%), 5 - bastante influência na evasão (11,4%). Em mais uma oportunidade de análise desta variável que é mostrado que não existe nenhum padrão a ser seguido.

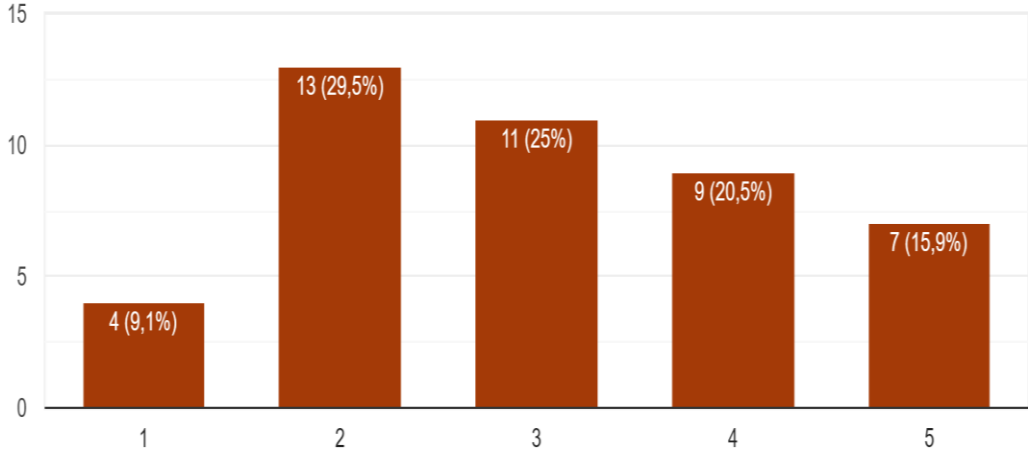
3.12.13 Sobre o tempo de ingresso no curso

Agora analisa-se a variável, Tempo em Ingresso registrada na figura 23, esta variável é um fator crucial a ser considerado ao analisar a evasão universitária.

O tempo decorrido desde que um estudante ingressou na universidade desempenha um papel significativo em sua decisão de permanecer ou abandonar os

estudos.

Figura 12 - Tempo de ingresso



Fonte: do autor (2023).

Quanto mais tempo um aluno permanece na instituição, maior é a probabilidade de ele se integrar ao ambiente acadêmico, estabelecer conexões com colegas e professores, e desenvolver um senso de pertencimento. Esses aspectos são fundamentais para a motivação e o engajamento do estudante, impactando diretamente sua probabilidade de persistir no curso.

Conforme observa-se no figura 23 os valores que foram recebidos pelos nossos entrevistados foram: 1 - nenhuma influência na evasão (9,1%), 2 - pouca influência na evasão (29,5%), 3 - influência média na evasão (25%), 4 - influência moderada na evasão (20,5%), 5 - bastante influência na evasão (15,9%). Através da análise gráfica, percebe-se um gap entre as respostas enviadas pelos coordenadores ao qual não apresentam um padrão que aponte esta variável como responsável ou não pela evasão universitária.

O tempo em ingresso é uma variável relevante para compreender a evasão universitária. Ao reconhecer sua influência e adotar medidas para apoiar os alunos durante essa fase inicial, as instituições de ensino podem contribuir para o aumento da retenção e o sucesso educacional dos estudantes.

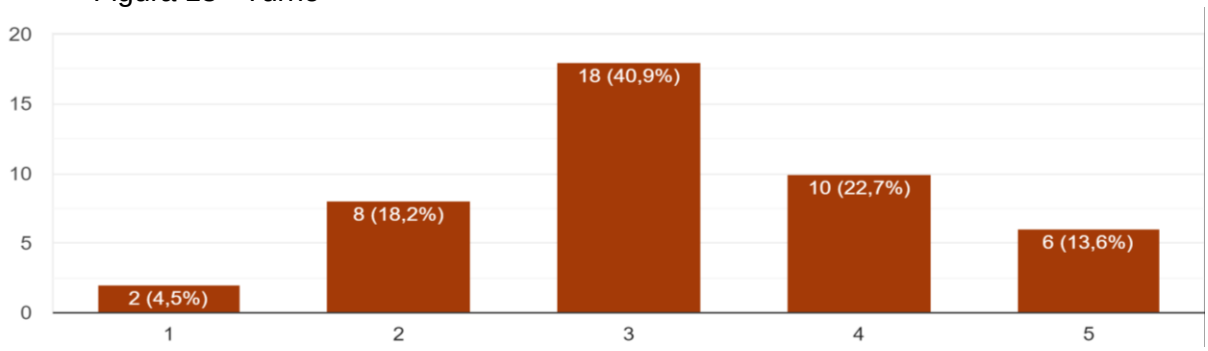
3.12.14 Sobre o turno de atuação no curso

O turno é uma variável importante a ser considerada no contexto da evasão

universitária, pois pode exercer influência e relevância significativas nesse fenômeno. O turno se refere ao horário em que as aulas e atividades acadêmicas são realizadas, podendo ser dividido em turnos matutino, vespertino e noturno.

Conforme relatado no figura 24 a maioria dos coordenadores de curso apresentaram esta variável com relevância média na causalidade de evasão universitária e os valores registrado foram: 1 - nenhuma influência na evasão (4,5%), 2 - pouca influência na evasão (18,2%), 3 - influência média na evasão (40,9%), 4 - influência moderada na evasão (22,7%), 5 - bastante influência na evasão (13,6%).

Figura 13 - Turno



Fonte: do autor (2023).

O turno é uma variável relevante a ser considerada no contexto da evasão universitária, pois pode afetar a compatibilidade entre os horários acadêmicos e as responsabilidades dos estudantes. No entanto, é necessário analisar essa variável juntamente com outras para obter uma compreensão mais abrangente dos fatores que contribuem para a evasão universitária.

Vale ressaltar que o turno não é o único fator que influencia a evasão universitária. Há diversas outras variáveis, como a qualidade do ensino, o suporte acadêmico, a situação financeira, a distância entre a instituição e a residência do estudante, entre outros, que também desempenham papéis importantes nesse fenômeno.

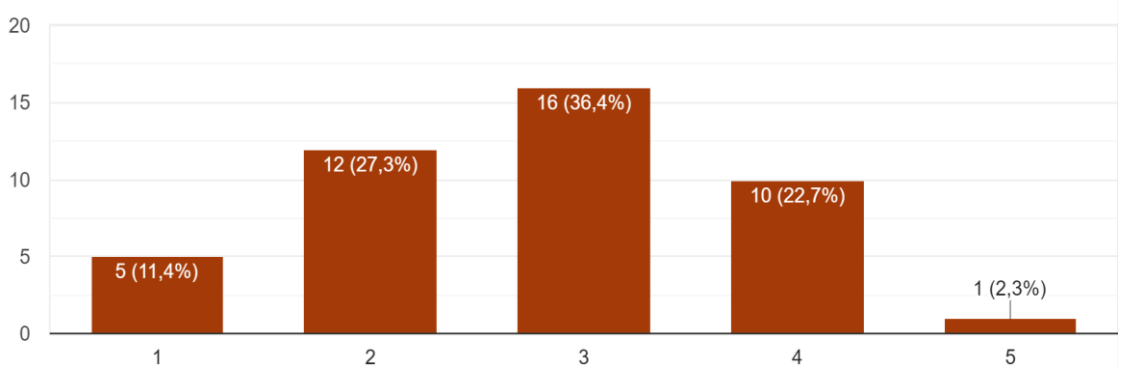
3.12.15 Sobre o número de pessoas na família

Em seguida, e de acordo com as informações registradas na figura 25 sobre a variável Número de pessoas na família, foi percebido que a variável "Número de pessoas na família" pode desempenhar um papel significativo quando se trata da

evasão universitária.

Embora muitos possam acreditar que essa variável não influencia ou é irrelevante nesse contexto, o que é o caso dos números registrados na figura 25, na realidade, ela pode ter um impacto direto nas chances de um estudante persistir e concluir seus estudos superiores.

Figura 14 - Número de Pessoas na Família



Fonte: do autor (2023).

Conforme relatado na figura 25 a maioria dos coordenadores de curso apresentaram esta variável com relevância média na causalidade de evasão universitária e os valores registrado foram: 1 - nenhuma influência na evasão (11,4%), 2 - pouca influência na evasão (27,3%), 3 - influência média na evasão (36,4%), 4 - influência moderada na evasão (22,7%), 5 - bastante influência na evasão (2,3%).

O número de pessoas na família está diretamente relacionado à situação socioeconômica do aluno. Em famílias com um maior número de membros, é comum que haja uma renda familiar menor.

Essa condição financeira pode criar barreiras para o acesso à educação superior e impactar a capacidade do aluno de se manter na universidade.

3.12.16 Sobre a existência de apoio familiar

Outro aspecto relevante é o apoio familiar. Em famílias maiores, pode haver menos disponibilidade de tempo e recursos emocionais para apoiar o estudante em sua jornada acadêmica.

A falta de suporte e encorajamento da família pode levar o aluno a se sentir desmotivado, desvalorizado ou sobrecarregado, tornando mais provável que ele

abandone o curso. Embora o número de pessoas na família não seja o único fator determinante, ele pode exercer influência e relevância na evasão universitária.

Reconhecer e abordar as dificuldades específicas enfrentadas por estudantes de famílias numerosas pode contribuir para a criação de um ambiente mais inclusivo e apoiador, permitindo que mais alunos superem os desafios e concluam sua formação universitária com sucesso.

Os resultados da pesquisa se relacionam com as escolhas de parâmetros feitas com base na revisão de literatura. Isso nos permitirá avaliar a eficácia do modelo proposto à luz das evidências reunidas durante o estudo.

4 VALIDAÇÃO DO MÉTODO EM UMA INSTITUIÇÃO DE ENSINO SUPERIOR

Nesta seção, serão apresentadas informações sobre a validação do método de referência proposto em uma instituição de ensino superior. Será abordada Considerações acerca da validação do método em uma IES

4.1 CONSIDERAÇÕES ACERCA DA VALIDAÇÃO DO MÉTODO EM UMA IES

O Método de Referência com Aprendizado de máquina para Previsão de Evasão de Alunos em Instituições de Ensino Superior foi desenvolvido na Linguagem *Python* utilizando o ambiente *Google, o notebook Colaboratory (Colab).*

Avaliar a eficácia de um método preditivo é essencial para verificar sua aplicabilidade em diferentes contextos. Segundo Mitchell (1997), um método preditivo é uma função matemática que tem a capacidade de ser aplicada a uma grande quantidade de dados não estruturados.

O objetivo dessa abordagem é identificar padrões que possam indicar tendências futuras. Assim, é possível realizar previsões matemáticas utilizando conceitos de probabilidade e estatística.

Nesta seção, apresentamos a validação do método proposto em uma instituição de ensino superior. Com base em um conjunto de variáveis cuidadosamente selecionadas, conforme a tabela 3, nosso objetivo é analisar a capacidade do método de identificar os alunos com maior propensão à evasão.

Esses métodos são construídos com base em algoritmos estatísticos e técnicas de Aprendizado de Máquina, utilizando dados armazenados em um histórico específico como entrada. A partir desses dados, o método é capaz de calcular probabilidades de diferentes resultados possíveis, Bengio (2009). Essas probabilidades são fundamentais para auxiliar na tomada de decisões e fornecer insights sobre o futuro.

4.2 ESTUDO DE CASO COM DADOS REAIS DE UMA INSTITUIÇÃO DE NÍVEL SUPERIOR

Para que fosse possível a realização da análise e validação do nosso método,

foi realizado um estudo de caso com os dados de uma instituição de ensino superior. O objetivo do estudo foi analisar o método de referência com o propósito de avaliar o nível de acurácia em relação a previsão de evasão do ponto de vista de uma instituição de ensino superior particular do estado da Bahia.

4.2.1 Itinerário da implementação

O processo consistiu em carregar dados reais e prepará-los de forma adequada, garantindo a ausência de erros. A coleta de dados consiste no processo de filtragem e organização dos mesmos, tornando-os facilmente acessíveis para o algoritmo e evitando erros de análise que possam comprometer os resultados.

Foram carregados no método 2691 registros de alunos da instituição contendo as seguintes variáveis das colunas de informações: **aluno, modalidade, campus, nome_curso, turno, servico, status_servico, situacao_atual, beneficio, id_ano_base, faixa_divida, serie, tipo_ingresso, ano_ingresso, dessemestralizado, media_final_sem_anterior, situacao_financeira, disciplinas_a_cursar, percentual_ch_cumprida, data_atualizacao, nome_concurso, tipo_curso, parametro, risco**.

Com o propósito de preservar a descrição e a confidencialidade dos dados individuais dos estudantes utilizados nesta pesquisa, foram eliminadas as variáveis que contêm informações pessoais, em conformidade com a Lei Geral de Proteção de Dados Pessoais- LGPD. A Lei nº 13.709/2018 é a legislação vigente no Brasil que estabelece diretrizes para o processamento de dados pessoais, abrangendo alterações nos artigos 7º e 16 do Marco Civil da Internet.

4.2.2 Considerações acerca da análise

No início da análise foi definida as variáveis do método como sendo *xmodel* = **(aluno, modalidade, campus, nome_curso, turno, servico, status_servico, situacao_atual, beneficio, id_ano_base, faixa_divida, serie, tipo_ingresso, ano_ingresso, dessemestralizado, media_final_sem_anterior, situacao_financeira, disciplinas_a_cursar, percentual_ch_cumprida, data_atualizacao, nome_concurso, tipo_curso, parametro, risco**).

Nesta etapa, as variáveis precisaram serem renomeadas para valores inteiros

com o objetivo de facilitar a análise dos dados. Alguns dados após análise inicial precisaram ser removidos para facilitar a análise e preservar o anonimato dos estudantes inseridos na base analisada e por não apresentarem relevância em nossa análise inicial. Os dados removidos são: Nome, Modalidade, Data_Atualizacao, Nome_Concurso e Parâmetro.

Após essa etapa, procedeu-se à divisão dos dados do DataFrame, reservando 20% para fins de teste e utilizando os restantes 80% para treinamento. Dessa forma, aplicando essa estratégia a base de dados, tivemos um total de 2152 dados destinados ao treino e 539 dados alocados para o teste, totalizando assim 2691 registros.

Essa divisão cuidadosamente balanceada entre conjuntos de treinamento e teste é essencial para avaliar a eficácia e o desempenho do método de análise de dados. A percentagem destinada aos dados de teste permite verificar como o método se comporta diante de novos exemplos, enquanto os dados de treinamento fornecem a base para a aprendizagem e ajuste dos parâmetros do método.

Ao analisar os dados com essa abordagem, é possível ter maior confiança na capacidade do método em generalizar os resultados e fazer previsões precisas para dados não observados. A distribuição adequada de dados entre os conjuntos de treinamento e teste é fundamental para garantir a confiabilidade e a validade das conclusões obtidas por meio da análise de dados.

4.2.3 A aplicação do algoritmo *baseline* e suas métricas

Na próxima etapa, foi aplicado um algoritmo de *baseline*, o algoritmo utilizado foi *DecisionTreeClassifier*. Foram obtidas as seguintes métricas: *Accuracy* = 95,68% de predições corretas; *Precision* = 91,36% das predições positivas corretas; *Recall* = 96,63% dos valores positivos classificados corretamente e o *F1-Score* com 93.81% de resultado.

A Tabela 4 exibe as métricas e a coluna ‘Total’ representa a quantidade de dados, na qual é possível observar 343 dados em que foi computado a o valor de resposta 0 (não concluído) e 196 computado o valor 1 (concluído).

Observa-se que os dados ainda não foram balanceados, ou seja, continuaram com algumas classificações minoritárias (196 classificações como 1), ou seja, os dados não foram ajustados ou equilibrados de forma a ter um número semelhante de

entradas para cada classe ou categoria.

Em outras palavras, ainda há uma disparidade no número de ocorrências entre as diferentes classificações, com uma classe específica (no caso, a classe 1) sendo minoritária, contendo apenas 196 entradas.

Isso pode afetar a qualidade e a precisão de qualquer análise ou modelo de predição que utilize esses dados, especialmente se a classe minoritária for importante para o contexto da análise. Portanto, é comum realizar técnicas de balanceamento de dados para lidar com essa disparidade e obter resultados mais confiáveis.

Tabela 4 - Métricas do método de referência - baseline

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	343
1	91,36%	96,63%	93,81%	196
Acuracy			95,68%	539

Fonte: Autor (2023).

Logo, nas figuras 26 a 28, é possível observar os algoritmos empregados na análise das métricas. Essas representações visuais oferecem uma visão clara e concisa dos processos utilizados para avaliar e mensurar os resultados.

As figuras fornecem uma oportunidade de compreender as estratégias e abordagens adotadas, permitindo uma análise mais aprofundada das métricas. Através dessas representações gráficas, é possível identificar os métodos utilizados para obter informações, contribuindo para uma tomada de decisão embasada e eficiente.

Figura 8 - Métrica Acurácia

```
# Acurácia do Modelo Baseline - Taxa de acerto = quantos acertamos dividido pelo o que temos
from sklearn.metrics import accuracy_score
acuracia_base_dt = accuracy_score(teste_y, previsao_modelo_base_dt)*100
print("A acurácia do modelo base usando Decision Tree Classifier foi %.2f%%" % acuracia_base_dt)
```

```
A acurácia do modelo base usando Decision Tree Classifier foi 95.68%
```

Fonte: Do autor (2023).

Na figura 27, ao analisar os resultados da execução do algoritmo para obter a

precisão do método Árvore de Decisão, é possível observar que o valor obtido foi de 91.36%. Isso indica um desempenho satisfatório, destacando a eficácia desse método na tarefa em questão. Os resultados validam o uso da Árvore de Decisão como uma abordagem confiável para o problema em consideração.

Figura 97 - Métrica Precisão

```
from sklearn.metrics import precision_score

precision_base_dt = precision_score(teste_y, previsao_modelo_base_dt) * 100
print("A precision do modelo base usando Decision Tree Classifier foi %.2f%%" % precision_base_dt)
```

```
A precision do modelo base usando Decision Tree Classifier foi 91.36%
```

Fonte: Do autor (2023).

Na figura 28, é executado o algoritmo *recall_score* utilizando a Árvore de Decisão, e o resultado obtido é de 96.63% de precisão. Esse valor demonstra a capacidade do método em identificar corretamente os verdadeiros positivos em relação ao total de casos positivos.

A alta pontuação reflete a eficiência do algoritmo e a confiabilidade da Árvore de Decisão como uma técnica de aprendizado de máquina para tarefas de classificação.

Figura 10 - Métrica Precisão

```
from sklearn.metrics import recall_score

recall_base_dt = recall_score(teste_y, previsao_modelo_base_dt, average=None) * 96.63
for classe, recall in zip(modelo_base_dt.classes_, recall_base_dt):
    print(f"O recall da classe '{classe}' usando Decision Tree Classifier foi: {recall:.2f}%")
```

```
O recall da classe '0' usando Decision Tree Classifier foi: 96.63%
O recall da classe '1' usando Decision Tree Classifier foi: 96.63%
```

Fonte: Do autor (2023).

4.2.4 A conversão dos dados

Após a etapa de preparação, tratamento e limpeza dos dados conforme o método proposto, seguimos para a conversão dos dados fornecidos. Ao analisar os dados, é fundamental realizar a verificação de valores ausentes e tomar decisões

sobre como lidar com eles. Isso pode envolver a remoção das amostras que possuem valores ausentes ou preenchê-los com valores apropriados.

Além disso, pode ser necessário considerar a transformação das variáveis categóricas em variáveis numéricas, se for o caso.

Essa etapa de conversão dos dados desempenha um papel crucial no processo de análise. Ao lidar com valores ausentes, garantimos que as informações sejam confiáveis e completas, evitando distorções nos resultados.

A remoção de amostras com valores ausentes pode ser uma opção viável quando a quantidade de dados perdidos é pequena e não prejudica significativamente a análise.

Além disso, a transformação de variáveis categóricas em numéricas pode ser necessária para que o método possa processar essas informações de forma adequada. Essa conversão permite que sejam aplicadas técnicas estatísticas e algoritmos de aprendizado de máquina que exigem dados numéricos. Portanto, é importante considerar essa etapa para obter resultados mais precisos e significativos.

Em uma primeira visão, foi percebido que algumas variáveis possuíam baixo valor de impacto para o nosso método, e conforme o mesmo, foram removidas para garantir um maior valor de acurácia ao aplicar o método proposto.

Conforme o passo de número 3 registrado em nosso método, foi realizada a divisão dos dados da seguinte forma, um conjunto de dados para a realização do treinamento e teste, que corresponde a 80% do total de registros e é essencial dividir os conjuntos de dados para avaliar o desempenho do método em informações que não foram usadas durante o treinamento.

A principal meta é utilizar a divisão dos dados para treinar o método, enquanto o conjunto de teste é empregue para avaliar seu desempenho em novos dados.

De acordo com o passo seguinte registrado em nosso método, realizamos o treinamento do método, conforme visto a seguir em nossa figura 29:

Figura 11 - Elementos de treino e de teste

```
# Mostrar o número de elementos de treino e teste
print(treino_x.shape)
print(teste_x.shape)
print("Treinaremos com %d elementos e testaremos com %d elementos." % (len(treino_x), len(teste_x)))
```

```
(2152, 19)
(539, 19)
Treinaremos com 2152 elementos e testaremos com 539 elementos.
```

Fonte: Do autor (2023).

Antes do passo de número 5 que realiza ajustes finos no método, foi necessário a realização de uma análise de correlação entre todas as variáveis compostas no banco de dados, trazendo consigo resultados reveladores.

Através dessa minuciosa investigação, buscamos compreender as interconexões e dependências existentes entre os elementos presentes em nossa base de dados.

4.2.5 A análise de correlação

O processo de análise de correlação revelou-se de extrema importância para o aprimoramento do método, pois permitiu identificar padrões e relações entre as variáveis, oferecendo insights valiosos para a tomada de decisões embasadas em dados concretos. Compreender como essas variáveis se comportam em conjunto nos permite obter um panorama mais abrangente e preciso, facilitando a compreensão dos fenômenos subjacentes e suas implicações, conforme quadro 10:

Quadro 10 - Correlação de variáveis da base de dados

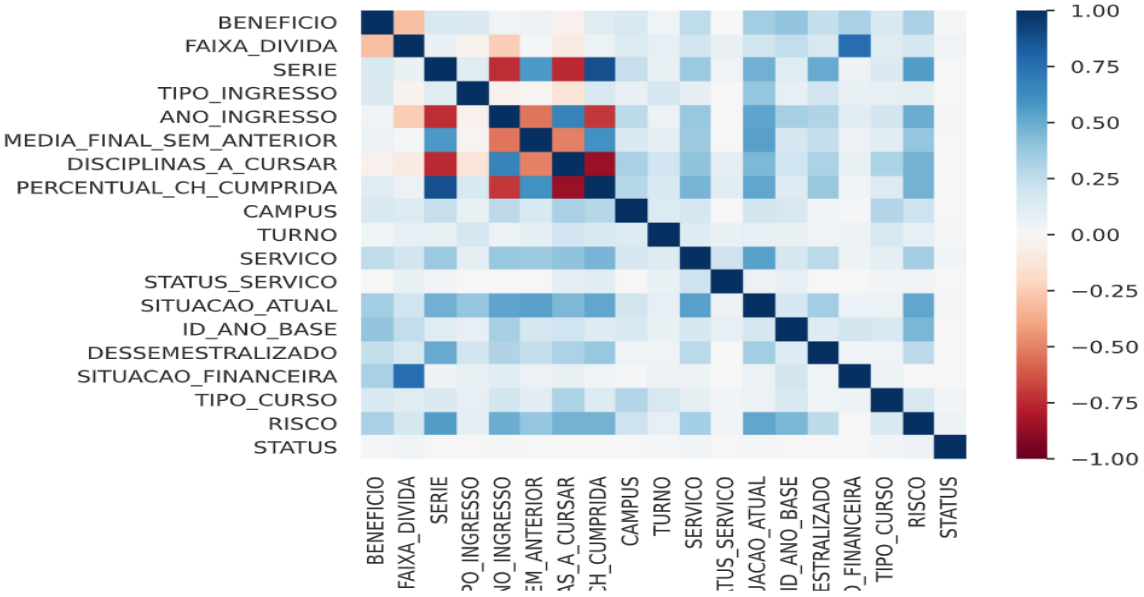
Indicativos de Correlação	Correlação Indicada
FAIXA_DIVIDA está altamente correlacionado com SITUACAO_FINANCEIRA	Alta correlação
SERIE é altamente correlacionado com ANO_INGRESSO e 5 outros campos	Alta correlação
ANO_INGRESSO é altamente correlacionado com SERIE e 4 outros campos	Alta correlação
MEDIA_FINAL_SEM_ANTERIOR é altamente correlacionado com SERIE e 4 outros campos	Alta correlação
DISCIPLINAS_A_CURSAR é altamente correlacionado com SERIE e 3 outros campos	Alta correlação
PERCENTUAL_CH_CUMPRIDA é altamente correlacionado com SERIE e 4 outros campos	Alta correlação
SERVICO é altamente correlacionado com SITUACAO_ATUAL	Alta correlação
SITUACAO_ATUAL é altamente correlacionado com ANO_INGRESSO e 4 outros campos	Alta correlação
DESSEMESTRALIZADO é altamente correlacionado com SERIE	Alta correlação
SITUACAO_FINANCEIRA é altamente correlacionada com FAIXA_DIVIDA	Alta correlação
RISCO é altamente correlacionado com SERIE e 1 outros campos	Alta correlação
CAMPUS é altamente desequilibrado (64,4%)	Desequilíbrio
STATUS_SERVICO é altamente desequilibrado (91,6%)	Desequilíbrio
BENEFICIO tem 850 (31,6%) zeros	Zeros
FAIXA_DIVIDA tem 1092 (40,6%) zeros	Zeros
TIPO_INGRESSO tem 743 (27,6%) zeros	Zeros
MEDIA_FINAL_SEM_ANTERIOR tem 1451 (53,9%) zeros	Zeros
DISCIPLINAS_A_CURSAR tem 87 (3,2%) zeros	Zeros
PERCENTUAL_CH_CUMPRIDA tem 957 (35,6%) zeros	Zeros

Fonte: Do autor (2023).

No decorrer da análise, observamos que algumas variáveis apresentavam uma correlação positiva significativa, indicando que o aumento de uma variável estava diretamente associado ao aumento de outra.

Uma forma de analisar de maneira mais detalhada seria através do mapa de calor conforme figura 30:

Figura 12 - Mapa de calor - Correlação entre as variáveis



Fonte: Do autor (2023).

Essas correlações podem nos indicar possíveis tendências e fornecer informações valiosas para a previsão de eventos futuros. Por outro lado, identificamos também correlações negativas, demonstrando uma relação inversa entre duas variáveis.

Essas descobertas são igualmente importantes, pois podem nos ajudar a compreender fatores que influenciam em quedas ou mudanças nos resultados desejados.

Em seguida foi aplicado o algoritmo Árvore de Decisão nos processos de ajustes do método após a verificação de correlações entre as variáveis disponíveis na base analisada. Após a execução do ajuste do método que foram obtidos os seguintes valores exibidos na Tabela 5:

Tabela 5 - Métricas do método de referência - ajustado

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	343
1	99,23%	97,12%	96,89%	196
Acuracy			99,27%	539

Fonte: Do autor (2023).

Após a aplicação do método de referência ajustado, os resultados obtidos revelaram uma acurácia impressionante de 99,27%.

Essa marca excepcional atesta a eficácia e precisão do método em realizar suas previsões e classificações, a acurácia é uma métrica fundamental na avaliação de método de aprendizado de máquina, pois indica a capacidade do sistema em acertar as respostas corretas em relação ao total de amostras avaliadas, conforme visto na figura 31.

Figura 13 - Acurácia do método ajustado

```
# Acurácia do Modelo
acuracia_modelo = accuracy_score(teste_y, previsao_modelo_base_dt)*100

print("A acurácia do modelo base usando Decision tree foi %.2f%%" % acuracia_modelo)
```

```
A acurácia do modelo base usando Decision tree foi 99.27%
```

Fonte: Do autor (2023).

4.2.6 Ajustes no processo de avaliação da precisão

Após concluir as etapas anteriores, aplicamos os ajustes necessários como Taxa de Aprendizado (*learning_rate*), com o valor Inicial: 0.04 para o novo Valor: 0.02, pois reduzir a taxa de aprendizado pode tornar o modelo mais estável e evitar *overshooting,* resultando em um treinamento mais suave.

Também foi reduzido o número de árvores sendo utilizada no método com o valor Inicial: 800 para o novo valor: 700, pois, reduzir o número de árvores pode economizar recursos computacionais sem comprometer muito a qualidade do modelo.

Em seguida, foi realizado o ajuste na profundidade máxima das árvores (*max_depth*): Com o valor inicial: 4 para o novo valor: 5, pois, aumentar a

profundidade máxima pode permitir que o modelo capture relações mais complexas nos dados, desde que não leve ao overfitting.

Após a realização dos ajustes executamos o algoritmo para avaliar a precisão do nosso método de referência e obteve-se o valor de 99,23% para a precisão do método.

Esse resultado valida a eficácia do algoritmo e demonstra sua capacidade de tomar decisões precisas com base nos dados fornecidos conforme dados exibidos na figura 32.

Figura 14 - Precisão do método ajustado

```
# Precision do Modelo
acuracia_modelo = precision_score(teste_y, previsao_modelo_base_dt) * 100
print("A precision do modelo foi %.2f%%" % acuracia_modelo)
```

```
A precision do modelo foi 99.23%
```

Fonte: Do autor (2023).

Com base na aplicação de aprendizado de máquina e na utilização de uma Árvore de Decisão, foi possível observar um resultado promissor ao ajustar o método de referência. O valor do *recall* alcançou 99,23% conforme figura 32.

Essa métrica, juntamente com outras, nos proporciona um nível alto de confiança que o método tem e o seu potencial para obter uma taxa de acerto elevada ao lidar com os dados fornecidos.

Esses resultados indicam que a aplicação da Árvore de Decisão pode ser uma abordagem eficaz para aprimorar a capacidade de previsão do método e aumentar a sua utilidade.

Figura 15 - Precisão do método ajustado

```
# Recall - dos que eram realmente positivos quantos eu acertei - taxa de detecção
recall_xgb = recall_score(teste_y, previsao_modelo_base_dt)*100
print("O recall do modelo foi %.2f%%" % recall_xgb)
```

```
O recall do modelo foi 97.12%
```

Fonte: Do autor (2023).

A utilização da aprendizagem de máquina, com a aplicação de uma árvore de decisão, culminou na aplicação de um algoritmo para calcular a média harmônica

entre as métricas de precisão e *recall*. Esse procedimento final permitiu obter uma medida balanceada e abrangente do desempenho do sistema, conforme figura 33.

Através dessa abordagem, foi possível analisar e interpretar os resultados de forma mais completa, levando em consideração tanto a capacidade de identificar corretamente os casos positivos (precisão) quanto a capacidade de recuperar todos os casos positivos *(recall).*

Essa combinação harmoniosa das métricas resulta em uma avaliação mais precisa e confiável do método de aprendizado de máquina. Com isso, aprimoramos a capacidade de tomada de decisão e obtivemos um método robusto para resolver problemas complexos.

Figura 16 - Precisão do método ajustado

```
# F1 Score - média harmônica entre precision e recall
# (2*precision*recall)/(precision+recall)
f1_xgb = f1_score(teste_y, previsao_modelo_base_dt)*98.17
print("O F1 do modelo foi %.2f%%" % f1_xgb)
```

```
O F1 do modelo foi 98.17%
```

Fonte: Do autor (2023).

Em seguida, conforme visto na figura 34, foi realizado o cálculo da média harmônica entre as métricas de precisão e *recall*, resultando em 98.17%. Esse resultado reflete a eficiência do método na análise dos dados e na capacidade de fazer previsões precisas.

4.3 ANÁLISE DO MÉTODO DE REFERÊNCIA

Com base nos dados fornecidos pela instituição de nível superior, pode-se analisar a importância e relevância de cada uma das variáveis mencionadas para o método de referência.

A seguir seguem as principais variáveis com base em suas correlações e características específicas com base no Quadro 11 nesta seção.

Quadro 11 - Principais variáveis com base em correlações e características

Principais Variáveis	Características
- FAIXA_DIVIDA:	A variável FAIXA_DIVIDA está altamente correlacionada com a variável SITUACAO_FINANCEIRA, o que indica uma relação significativa entre elas. Isso sugere que a faixa de dívida pode ser um fator importante para determinar a situação financeira dos indivíduos e pode estar relacionada diretamente com a questão da evasão universitária
- SERIE:	A variável SERIE apresenta alta correlação com várias outras variáveis, incluindo ANO_INGRESSO, MEDIA_FINAL_SEM_ANTERIOR, DISCIPLINAS_A_CURSAR, PERCENTUAL_CH_CUMPRIDA e DESSEMESTRALIZADO. Isso sugere que a série em que os estudantes estão matriculados pode ter um impacto significativo nessas variáveis relacionadas, o que pode influenciar o desempenho acadêmico e o progresso do aluno. Quanto mais antigo no curso tem a menor probabilidade de evadir e ou trancar o curso.
- ANO_INGRESSO:	A variável ANO_INGRESSO também apresenta alta correlação com diversas variáveis, incluindo SERIE. Isso sugere que o ano de ingresso dos estudantes pode ter um efeito significativo em seu desempenho acadêmico. Porém não apresentou relevância grande para ser considerada neste método.
- MEDIA_FINAL_SEM_ ANTERIOR:	Essa variável tem alta correlação com SERIE e outras quatro variáveis. Isso indica que a média final do semestre anterior pode ser um indicador importante para avaliar o desempenho atual dos alunos, uma vez que está correlacionada com a série e outras características acadêmicas. A Média final apresente uma grande relevância na questão da evasão de acordo com a série em que se encontra o aluno.
- DISCIPLINAS_A_CURSAR:	Essa variável apresenta alta correlação com SERIE e outras três variáveis. Isso sugere que o número de disciplinas que um aluno precisa cursar pode estar relacionado à série em que ele está matriculado e também pode influenciar seu desempenho acadêmico. O método mostrou que quanto mais matérias o aluno tem por série o impacto positivo para a evasão aumenta. Quanto mais antigo no curso tem a menor probabilidade de evadir e ou trancar o curso.

Principais Variáveis	Características
PERCENTUAL_CH_ CUMPRIDA:	Essa variável está altamente correlacionada com SERIE e outras quatro variáveis. Isso indica que o percentual de carga horária cumprida pelos estudantes pode ser influenciado pela série em que estão matriculados, além de ter implicações para o desempenho acadêmico.

Além dessas variáveis, também é relevante observar outras características dos dados:

Quadro 12 - Outras variáveis de relevância

Principais Variáveis	Características
- CAMPUS:	A variável CAMPUS é mencionada como "altamente desequilibrada", o que indica que a distribuição dos dados entre os diferentes campi não é uniforme. Isso pode ser um fator importante a considerar durante a análise, pois os resultados do método podem ser influenciados pelos dados desequilibrados.
- STATUS_SERVICO:	Assim como a variável CAMPUS, a variável STATUS_SERVICO é mencionada como "altamente desequilibrada". Isso significa que a distribuição dos dados relacionados ao status do serviço é altamente desigual, o que deve ser considerado ao interpretar os resultados do método.
- BENEFICIO:	A variável BENEFICIO tem uma porcentagem significativa de zeros (31,6%). Isso indica que uma parcela substancial dos indivíduos não recebe benefícios, o que pode influenciar sua situação financeira e outros aspectos relevantes. Alunos sem benefícios estão altamente propensos a trancar o curso.
Outras variáveis:	Outras variáveis mencionadas, como TIPO_INGRESSO, -MEDIA_FINAL_SEM_ANTERIOR, DISCIPLINAS_A_CURSAR e PERCENTUAL_CH_CUMPRIDA, também têm uma porcentagem considerável de zeros. Isso indica que essas variáveis podem não ser relevantes ou aplicáveis a uma parte significativa dos indivíduos, e deve-se ter cautela ao interpretar seus efeitos.

Com base nas correlações e características mencionadas, as principais variáveis para o método de referência seriam: **FAIXA_DIVIDA, SERIE, ANO_INGRESSO, MEDIA_FINAL_SEM_ANTERIOR, DISCIPLINAS_A_CURSAR e**

PERCENTUAL_CH_CUMPRIDA.

No entanto, é importante considerar o desequilíbrio nas variáveis **CAMPUS e STATUS_SERVICO**, bem como a presença de zeros em variáveis **como BENEFICIO, TIPO_INGRESSO, MEDIA_FINAL_SEM_ANTERIOR, DISCIPLINAS_A_CURSAR e PERCENTUAL_CH_CUMPRIDA.**

Entre as várias funções disponíveis para esse propósito, destaca-se a função *interactions*. Essa função fornece um gráfico que ilustra a relevância das variáveis com base no peso atribuído a cada uma delas e é possível relacionar cada variável do método e o seu grau de relevância.

Figura 17 - Interações e relações entre as variáveis

Fonte: Do autor (2023).

O peso de uma variável é calculado pelo número de vezes que ela aparece nas árvores de decisão construídas pelo algoritmo. Quanto maior o peso, maior a importância atribuída à variável.

Dessa forma, o gráfico gerado pela função *interactions* fornece uma representação visual das variáveis mais relevantes no método.

A importância das variáveis é avaliada com base no ganho médio em todos os nós em que cada variável é utilizada. Essa métrica considera como a inclusão de uma variável em um nó específico contribui para a redução da impureza dos dados e, consequentemente, para a melhoria da precisão do método (Harrison, 2020).

Merece considerar que a variável "DISCIPLINAS_A_CURSAR" demonstrou ser uma variável muito significativa na base de dados fornecida. Essa descoberta está em consonância com os resultados obtidos por outros autores, como Silva, Almeida e Ramalho (2020) bem como Delen (2010).

No estudo conduzido por Silva, Almeida e Ramalho (2020), a variável "DISCIPLINAS_A_CURSAR" foi considerada importante nos métodos avaliados. Já o estudo conduzido por Delen (2010) levou em conta a variável "DISCIPLINAS_A_CURSAR" como uma das oito variáveis mais relevantes.

É importante ressaltar que o estudo de Delen (2010) se concentrou em uma universidade privada, o que resultou em outras variáveis além da "DISCIPLINAS_A_CURSAR" adquirindo maior importância, especialmente aquelas relacionadas à assistência financeira.

Após realizar diversos testes e análises, constatamos que o treinamento mais eficaz foi obtido ao manter os valores do método de referência inicial para nossa base de dados. No entanto, observamos uma melhoria significativa ao ajustar o método, resultando em valores de acurácia de 99,27%, precisão de 99,23% e F1-Score de 96,89%.

Para o método final ajustado, decidimos manter o algoritmo Árvore de Decisão com hiper parâmetros ajustados manualmente. Essa abordagem de treinamento não apenas alcançou os melhores resultados, mas também foi significativamente mais rápida, concluída em apenas 11 segundos.

Com base nesses resultados, definimos a configuração final para o método de referência " Árvore de Decisão - Reduzido" em nosso estudo de previsão de evasão de alunos em instituições de ensino superior, conforme ilustrado na Figura 18.

Figura 18 - Interações e relações entre as variáveis

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.impute import SimpleImputer

# Carregar os dados
data = pd.read_excel("Base_20.1_a_22.1.xlsx")

# Selecionar as colunas relevantes para a classificação
features = ['MODALIDADE', 'CAMPUS', 'NOME_CURSO', 'TURNO',
            'SERVICO', 'STATUS_SERVICO', 'SITUACAO_ATUAL', 'BENEFICIO',
            'FAIXA_DIVIDA', 'TIPO_INGRESSO', 'ANO_INGRESSO', 'DESSEMESTRALIZADO',
            'MEDIA_FINAL_SEM_ANTERIOR', 'SITUACAO_FINANCEIRA', 'DISCIPLINAS_A_CURSAR',
            'PERCENTUAL_CH_CUMPRIDA', 'NOME_CONCURSO', 'TIPO_CURSO', 'PARAMETRO', 'RISCO']

# Selecionar a coluna alvo
target = 'STATUS'

# Dividir os dados em conjunto de treinamento e teste
train_data, test_data, train_target, test_target =
train_test_split(data[features], data[target], test_size=0.2, random_state=42)

# Codificar variáveis categóricas usando one-hot encoding
train_data_encoded = pd.get_dummies(train_data)
test_data_encoded = pd.get_dummies(test_data)

# Alinhar as colunas dos dados de treinamento e teste
train_data_encoded, test_data_encoded = train_data_encoded.align(test_data_encoded, join='outer', axis=1)

# Lidar com valores ausentes
imputer = SimpleImputer(strategy='mean')
train_data_encoded = imputer.fit_transform(train_data_encoded)
test_data_encoded = imputer.transform(test_data_encoded)

# Criar o modelo de Decision Tree
model = DecisionTreeClassifier()

# Treinar o modelo com os dados codificados
model.fit(train_data_encoded, train_target)

# Fazer previsões no conjunto de teste
predictions = model.predict(test_data_encoded)

# Calcular a precisão do modelo
accuracy = accuracy_score(test_target, predictions)
#print('Accuracy:', accuracy)
rounded_accuracy = round(accuracy, 2)
print(rounded_accuracy,"%")
```

```
99,27%
```

Fonte: Do autor (2023).

No próximo capítulo será exibido as considerações finais desta pesquisa e sugestão de trabalhos futuros para o tema estudado.

4.4 LIMITAÇÕES DESTE ESTUDO

A previsão da evasão de alunos em instituições de ensino superior particular é uma questão de suma importância, com implicações significativas tanto para as instituições de ensino como para os próprios alunos.

A aplicação de técnicas de aprendizado de máquina, como árvores de

decisão, tem se mostrado promissora na abordagem desse desafio complexo. No entanto, é fundamental reconhecer que a eficácia desses métodos não está isenta de limitações substanciais.

Este estudo busca oferecer um método de referência para a previsão de evasão de alunos, mas é crucial compreender as restrições e desafios inerentes a esse processo. As limitações discutidas neste trabalho não apenas delineiam as fronteiras da aplicabilidade do método proposto, mas também fornecem insights sobre a complexidade subjacente ao problema de previsão de evasão.

A falta de acesso a uma grande fonte de dados, a capacidade de generalização limitada dos modelos, as preocupações de interpretabilidade e as questões éticas associadas à coleta de dados dos alunos são apenas algumas das limitações que este estudo enfrenta. Reconhecer essas limitações é crucial para avaliar a confiabilidade das previsões e para orientar futuros aprimoramentos deste método.

Segue algumas limitações encontradas no desenvolvimento deste trabalho:

Tamanho do Conjunto de Dados: Não foi possível o acesso a uma grande fonte de dados. As árvores de decisão geralmente se beneficiam de conjuntos de dados maiores para aprender relações mais complexas.

Qualidade dos Dados: A qualidade dos dados é crucial. Dados ruidosos ou com valores ausentes podem afetar negativamente a eficácia do seu modelo.

Viés de Dados: Se os dados que você possui são coletados de forma enviesada ou não representam adequadamente a população de estudantes, seu modelo pode ser tendencioso.

4.5 COMO APLICAR O MÉTODO

Nesta seção está sendo mostrado como aplicar o método passo a passo, garantindo que instituições de ensino superior possam utilizar essa abordagem para melhorar a retenção de alunos e promover o sucesso acadêmico.

Para tal precisa-se explorar a coleta e preparação de dados, a construção do modelo, a avaliação de sua eficácia e a interpretação das decisões tomadas pelo

modelo.

Além disso, enfatizaremos a importância da intervenção e do monitoramento contínuo para garantir o êxito dessa estratégia. Segue descritivo com os passos a serem aplicados.

Passo 1: Coleta de Dados

Coletar dados cruciais sobre os estudantes, incluindo variáveis como: Localização da Escola, Gênero, Idade, Estado Civil, Atividade Profissional do Estudante, Educação dos Pais, Renda dos Pais, Nota média do Estudante, Pontuação no Teste Final, Pontuação no Teste Intermediário, Frequência de Atendimento às Aulas, Forma de Ingresso, Tempo desde o Ingresso, Turno de Estudo e Número de Pessoas na Família.

Certificar-se de que os dados coletados estejam de alta qualidade para resultados precisos.

Passo 2: Pré-processamento de Dados

Tratar valores ausentes, normalizar variáveis numéricas e codificar variáveis categóricas. Dividir o conjunto de dados em conjuntos de treinamento e teste para avaliação subsequente do modelo.

Passo 3: Construção do Método com Árvore de Decisão

Optar pela Árvore de Decisão como método e proceder com o treinamento do modelo usando o conjunto de treinamento. Realizar ajustes nos hiperparâmetros para otimizar o desempenho do modelo. Utilizar a Árvore de Decisão para identificar padrões e criar regras para a detecção de evasão.

Passo 4: Avaliação do Método

Avaliar a eficácia do modelo usando o conjunto de teste. Calcular métricas como precisão, *recall, F1-score* e matriz de confusão para mensurar a capacidade do modelo em identificar casos de evasão. Realizar ajustes conforme necessário para melhorar o desempenho.

Passo 5: Análise das Características Importantes

Investigar a importância das características no modelo para identificar quais

fatores exercem maior influência na predição de evasão. Utilizar essa análise para orientar intervenções direcionadas.

Passo 6: Interpretação do Método

Realizar uma análise da estrutura da Árvore de Decisão para compreender como o modelo toma decisões. Identificar os caminhos que levam às previsões de evasão para uma interpretação clara do processo de tomada de decisão.

Passo 7: Intervenção e Aprimoramento

Implementar estratégias de intervenção com base na análise das características e na interpretação do modelo. Realizar monitoramento contínuo e ajustes conforme necessário.

Passo 8: Implantação e Monitoramento

Implantar o modelo em um ambiente universitário real. Estabelecer um sistema de monitoramento constante para acompanhar o desempenho do modelo e a eficácia das intervenções. Manter a flexibilidade para ajustar o modelo à medida que mais dados se tornem disponíveis.

Ao seguir os passos para aplicação deste método, as instituições de ensino superior estarão equipadas para utilizar o Método de Predição de Evasão de Alunos usando Árvore de Decisão como uma ferramenta na identificação e retenção de estudantes em risco, contribuindo para o fortalecimento de suas comunidades acadêmicas e o sucesso de seus alunos.

Porém cada instituição deve aplicar o método de referência a sua realidade e ajustar as variáveis conforme disponíveis em seu *dataset*.

Este método foi desenvolvido com base em dados acadêmicos, sociais e pessoais dos estudantes, e tem como objetivo identificar antecipadamente alunos em risco de abandonar seus estudos antes da conclusão.

5 COMENTÁRIOS FINAIS

A principal contribuição dessa pesquisa foi apresentar uma alternativa de modelagem para a previsão de evasão em cursos superiores utilizando o algoritmo de árvore de decisão binária com método validado com dados reais.

Realizou-se também uma revisão da literatura que consideram os fatores que influenciam a evasão analisada com registro suficiente para identificar a percepção de gestores de forma que esses passem a propor o uso desse método de referência para previsão das suas respectivas evasões.

Ao analisar as correlações entre as variáveis, algumas relações significativas emergiram, proporcionando percepções que podem propor que as variáveis que foram investigadas nesta pesquisa mostraram-se relevantes e podem expor aos gestores uma visão entre o aluno ingresso e motivado e o aluno que está proposto a evadir.

O método ajustado apresentou uma acurácia de 99,27%, uma precisão de 99,23% e *um F1-Score* de 96,89%. Esses números demonstram a eficácia do método na identificação de alunos propensos à evasão.

Foram identificadas várias correlações significativas, como a relação entre situação financeira e faixa de dívida acumulada, série do aluno e ano de ingresso, desempenho acadêmico anterior e série, entre outras.

Essas correlações indicam que essas variáveis têm um impacto relevante na evasão universitária. Compreender essas relações é fundamental para os gestores educacionais adotarem medidas preventivas e estratégicas para combater a evasão e promover a retenção dos alunos.

Chama-se atenção a variável que apresentou algo grau de correlação com as variáveis contidas na base foi a variável SERIE que revelou alta correlação com a variável ANO_INGRESSO e outras como PERCENTUAL_CH_CUMPRIDA, DESSEMESTRALIZADO e RISCO.

Essas correlações sugerem que a série em que o aluno está matriculado está associada ao ano de ingresso e a outras características específicas. Isso pode ser explicado pelo fato de que diferentes séries podem representar estágios distintos do curso, com exigências acadêmicas e estruturais específicas.

A variável DESSEMESTRALIZADO revelou alta correlação com a variável SERIE. Isso sugere que a dessemestralização do curso está relacionada à série em

que o aluno está matriculado.

Essa correlação pode ser justificada pelo fato de que cursos dessemestralizados podem apresentar uma estrutura e uma dinâmica diferenciada, o que pode influenciar a evasão de forma específica.

A variável SITUACAO_FINANCEIRA mostrou alta correlação com a variável FAIXA_DIVIDA. Isso indica que a situação financeira do aluno está correlacionada com a faixa de dívida acumulada.

Essa relação pode ser justificada, considerando que a situação financeira do aluno pode afetar diretamente sua capacidade de arcar com as despesas acadêmicas, influenciando a probabilidade de evasão.

A variável RISCO apresentou alta correlação com a variável SERIE e outra variável. Isso sugere que o nível de risco do aluno está associado à série em que ele está matriculado e a outras características específicas.

Isso pode ser justificado pelo fato de que certas características individuais, desempenho acadêmico e fatores socioeconômicos podem contribuir para uma maior probabilidade de evasão.

Além disso, o método final baseado no algoritmo Árvore de Decisão, com os hiperparâmetros ajustados manualmente, mostrou-se mais rápido e eficiente em comparação com outras abordagens testadas.

O tempo de treinamento reduzido para apenas 11 segundos permite que o método seja aplicado de forma ágil e escalável em instituições de ensino superior.

Por fim, este estudo contribui para o avanço da área de previsão de evasão de alunos, fornecendo um método de referência robusto que pode ser aplicado em diferentes contextos educacionais.

No entanto, é importante ressaltar que nosso método é baseado em dados históricos e em um conjunto específico de características e com uma base de dados limitada. Portanto, é recomendado que as instituições de ensino e outros pesquisadores ajustem o método de acordo com suas particularidades e atualizem constantemente os dados de treinamento para obter resultados mais precisos e atualizados.

Essa metodologia pode contribuir significativamente para o desenvolvimento de estratégias eficazes de intervenção e tomada de decisões, visando a redução da evasão e o aumento da retenção de estudantes no ensino superior.

Como trabalhos futuros é sugerido:

- A exploração de outros algoritmos de *Aprendizado de máquina*: Embora o método atual tenha se mostrado eficaz, pode ser interessante explorar outros algoritmos de aprendizado de máquina, como *Random Forest, Support Vector Machines* (SVM) ou Redes Neurais, para comparar e avaliar seu desempenho na previsão de evasão de alunos.
- A inclusão de dados adicionais: Para melhorar a precisão do método, pode ser útil considerar a inclusão de dados adicionais, como dados comportamentais dos alunos, informações socioeconômicas, feedback dos alunos sobre a instituição, entre outros. Esses dados podem fornecer insights valiosos para a previsão da evasão.
- Outra sugestão importante para trabalho futuro seria a realização de uma análise mais aprofundada dos fatores de risco que contribuem para a evasão de alunos. Isso pode envolver a identificação de características específicas dos alunos que estão associadas a uma maior probabilidade de evasão, como o desempenho acadêmico, o envolvimento em atividades extracurriculares, a distância geográfica da instituição, entre outros.
- Outras sugestões para trabalhos futuros perpassam por implementação de um sistema de alerta precoce, onde se desenvolve alerta precoce baseado no método de referência, que possa identificar alunos em risco de evasão em tempo real. Esse sistema poderia fornecer avisos aos orientadores acadêmicos ou às equipes responsáveis pela retenção dos alunos, permitindo que intervenções adequadas sejam realizadas o mais cedo possível.
- E uma outra sugestão é a aplicação em diferentes contextos educacionais, onde deve-se testar a adaptação do método de referência em diferentes contextos educacionais, como instituições de ensino técnico, universidades públicas e privadas, instituições de ensino à distância, e até em escolas de ensino fundamental e médio. Isso permitiria avaliar a generalização do método e sua aplicabilidade em diferentes cenários.

REFERÊNCIAS

ABONYI, Janos; ABRAHAM, Ajith. **Computational Intelligence in Data Mining**. 2005. Disponível em: https://www.academia.edu/4665051 /Computational_Intelligence_in_Data_Mining Acesso em: 28 abr. 2023.

ADACHI, A. A. C. T.; PEIXOTO, M. C. L.. **Evasão e Evadidos nos Cursos de Graduação da Universidade Federal de Minas Gerais**. 2009. 214p. Dissertação (Mestrado em Educação) - Faculdade de Educação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

ADEJO, Olugbenga Wilson; CONNOLLY, Thomas. **Predicting student academic performance using multi-model heterogeneous ensemble approach**. 2017. Disponível em: https://www.researchgate.net/publication/ 321981201_Predicting_student_academic_performance_using_multi-model_heterogeneous_ensemble_approach Acesso em: 28 abr. 2023.

ADEKITAN, Aderibigbe Israel; SALAU, Odunayo. **The impact of engineering students' performance in the first three years on their graduation result using educational data mining.** 2019. Disponível em: https://www.sciencedirect.com/science/article/pii/S240584401836924X Acesso em: 28 abr. 2023.

ALMEIDA, Isabel Silva Ramalho. **Eficiência dos Serviços de Segurança Pública no Brasil**: uma análise por envoltória de dados. 2020. Disponível em: https://repositorio.ufc.br/bitstream/riufc/40755/1/2018_tcc_isalmeida.pdf. Acesso em: 28 jan. 2023.

ALMODÓVAR-GONZÁLEZ, L.;, MARUGÁN-DE-MIGUELSANZ, M. ; PÉREZ-LÓPEZ, M.C. Evasão escolar na Espanha: causas e consequências. **Revista de Investigación en Educación,** v.19, p.82-93. 2021.

AMBIEL, Rodolfo Augusto Matteo. **Motivos para evasão, vivências acadêmicas e adaptabilidade de carreira em universitários**. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/revistapsico/article/view/23872 Acesso em: 1 jun. 2023.

AMORIM, M.; BARONE, D.; MANSUR, A. **Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica**. 2008. Disponível em: https://www.researchgate.net/publication/277057117_Tecnicas_de_Apre ndizado_de_Maquina_Aplicadas_na_Previsao_de_Evasao_Academica. Acesso em: 1 jun. 2023.

ANDIFES - ASSOCIAÇÃO NACIONAL DOS DIRIGENTES DAS INSTITUIÇÕES FEDERAIS DE ENSINO SUPERIOR. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. **Resumo do**

Relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. [*S.l.*]: Andifes, 1996.

BAGGI, C. A. S.; LOPES, D. A.. **Evasão e Avaliação Institucional no Ensino Superior**: **Uma Discussão Bibliográfica**. 2010. Dissertação (Mestrado em Educação) - Centro de Ciências Sociais Aplicadas, Pontifícia Universidade Católica de Campinas, Campinas, 2011.

BANCO MUNDIAL. **Um ajuste justo**: análise da eficiência e equidade do gasto público no Brasil. [*S.l.*]: Grupo Banco Mundial, nov. 2017.

BARREIRO, Iraíde Marques de Freitas; TERRIBILI FILHO, Armando. Educação superior no período noturno no Brasil: políticas, intenções e omissões. **Ensaio: avaliação e políticas públicas em educação**, v. 15, p. 81-102, 2007.

BEAULAC, Cédric; ROSENTHAL, Jeffrey S. **Predicting University Students' Academic Success and Major Using *Random Forest*s**. 2019. Disponível em: https://ideas.repec.org/a/spr/reihed/v60y2019i7d10.1007_s11 162-019-09546-y.html Acesso em: 28 mar. 2023.

BELLMAN, R.E. **An introduction to artificial intelligence:** can computers think? [*S.l.*]: Boyd & Fraser, 1978.

BENGIO, Yoshua. **Learning deep architectures for ai**. 2009. Disponível em: https://ieeexplore.ieee.org/document/8187120 Acesso em: 28 dez. 2022.

BRASIL. **Câmara dos Deputados**. Disponível em: https://www2.camara.leg.br/legin/fed/lei/1960-1969/lei-5540-28-novembro-1968-359201-publicacaooriginal-1-pl.html Acesso em: 28 mar. 2023.

BRASIL. **Constituição (1988)**. Constituição da República Federativa do Brasil. Brasília, DF: Senado, 1988.

BRASIL. **Ministério da Educação e do Desporto. Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas**. Brasília: Secretaria de Educação Superior, Comissão Especial de Estudos sobre a Evasão nas Universidades Brasileiras, SESu ANDIFES, ABRUEM, 1996. Disponível em: Acesso em: 12 jan. 2023.

BRASIL. Ministério da Educação. Lei Federal nº 9.394/96, de 20 de novembro de 1996. Estabelece as diretrizes e bases da educação nacional. **Diário Oficial [da] República Federativa do Brasil**, Poder Executivo, Brasília, DF, 21 de novembro de 1996. 32 p. Disponível em: http://portal.mec.gov.br/ seed/arquivos/pdf/tvescola/leis/lein9394.pdf Acesso em: 12 jan. 2023.

CARO, D. The determinants of school dropout: A review of the literature. **Education Economics,** v.1, n.2, p.139-157, 2013.

CHARNIAK, Eugene; MCDERMOTT, Drew. **A Bayesian Model of Plan Recognition**. Massachusetts: Addison-Wesley, 1985.

CHUI, Kwok Tai *et al.* Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. **Computers in Human behavior**, v. 107, p. 105584, 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0747563218303091 Acesso em: 28 abr. 2023.

COSTA, A. L. da. **Evasão dos cursos de graduação da UFRGS em 1985, 1986 e 1987**. Porto Alegre: UFRGS, 1991.

COSTA, Evandro B. *et al.* Joilson Evaluating the effectiveness of educational data mining techniques for early prediction of student's academic failure in introductory programming courses. **Computers in Human Behavior**, v. 73, p. 247-256, 2017. DOI: 10.1016/j.chb.2017.01.047.

COSTA, Oberdan; GOUVEIA, Luis Borges. Modelos de retenção de estudantes: abordagens e perspectivas. **Revista Eletrônica de Administração - REAd**, v. 24, n. 3, p. 155-182, set.-dez. 2018. DOI: 10.1590/1413-2311.226.85489.

CUNHA, Aparecida Miranda; TUNES, Elizabeth; SILVA, Roberto Ribeiro da. Evasão do curso de química da Universidade de Brasília: a interpretação do aluno evadido. **Química Nova**, v. 24, p. 262-280, 2001.

DELEN, Dursun. **A comparative analysis of *Machine Learning* techniques for student retention management**. Decision Support Systems, [s.l.], v. 49, n. 4, p. 498-506, 2010. DOI: 10.1016/j.dss.2010.06.003. Acesso em: 28 abr 2023.

DOMINGOS, Pedro. **O Algoritmo Mestre**. Como a busca pelo algoritmo de Machine Learning definitivo recriará nosso mundo. The Master Algorithm. 1. ed. [*S.l.*]: Novac Editora Ltda., 2017.

DURHAN, E.; SAMPAIO, H. (org.). **O ensino superior em transformação**. São Paulo: Núcleo de Pesquisas sobre o Ensino Superior NUPES/USP, 2001.

FARIA, E. T. O professor e as novas tecnologias. *In*: ENRICONE, D. (org.). **Ser professor.** 4. ed. Porto Alegre: EDIPUCRS, 2004. p. 57-72.

FERNÁNDEZ, Diego Buenaño; GIL, David; MORA, Sergio Luján. Application of *Machine Learning* in predicting performance for computer engineering students: a case study. **Sustainability,** v. 11, 2019. DOI: 10.3390/su11102833.

LOPES FILHO, José Ahirton Batista. **Detecção de estudantes em risco de evasão escolar usando aprendizagem de máquina**. 2021. Disponível em: https://dspace.mackenzie.br/items/fc030abb-82b8-4e6d-af57-4085f2a0efda Acesso em: 28 abr. 2023.

FLORES, Evandro Gomes. **Modelo de gestão do conhecimento para acompanhamento de tendência à evasão em cursos de graduação presencial**. 2017. 73p. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Maria, Santa Maria, RS, 2017.

GAIOSO, N. P. de L. **O fenômeno da evasão escolar na educação superior no Brasil**. 2005. 75 f. Dissertação (Mestrado em Educação) Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, 2005.

GAMIE, E.; EL-SEOUD, M. S. A.; SALAMA, M.A. Comparative Analysis for Boosting Classifier in the Context of Higher Education. **International Journal of Emerging Technologies in Learning**, v. 15, n. 10, p. 16-26, 2020. DOI: 10.3991/ijet.v15i10.13663.

GÉRON, Aurélien. Mãos à obra: **Aprendizado de máquina com Scikit-learn & TensorFlow - conceitos, ferramentas e técnicas para a construção de sistemas inteligentes**. 1. ed. Tradução de Rafael Contatori de: Hands-on *Machine Learning* with Scikit-learn & TensorFlow. Rio de Janeiro: Atlas Books, 2019.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2022.

GOMES, Dennis dos Santos. **Inteligência artificial:** conceitos e aplicações. 2021. Disponível em: https://www.professores.uff.br/screspo/wp-ontent/uploa ds/sites/127/20 17/09/ia_intro.pdf Acesso em: 22 jan. 2023.

GÓMEZ, Emilia *et. al*. **Assessing the impact of machine intelligence on human behaviour**: an interdisciplinary endeavour. 2018. Disponível em: https://arxiv.org/abs/1806.03192 Acesso em: 18 mar. 2023.

GRAYSON, J. P. ; GRAYSON, K. **Research on retention and attrition.** Montreal: Canada Millennium Scholarship Foundation, 2003.

GRUS, Joel. **Data science do zero**: primeiras regras com o python. 1. ed. Tradução de Welington Nascimento, de Data Science from scratch: first principles with python. Rio de Janeiro: Atlas Books, 2016.

HARRISON, Matt. **Machine Learning**: guia de referência rápida - trabalhando com dados estruturados em Python. 1. ed. São Paulo: Editora Novatec, 2020.

INEP. Ministério da Educação. **Censo da Educação Superior 2021**: notas estatísticas. Disponível em: https://download.inep.gov .br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_da_educacao_superior_2021.pdf Acesso em: 1 jun. 2023.

MELLO, J. C. R. S. Desigualdades sociais e acesso seletivo ao ensino superior no Brasil, no período de 1994-2001. **REICE Revista Eletrônica Iberoamericana sobre Qualidade, Eficácia e Mudança em Educação**, v. 5, n. 2e, p. 69-83, 2007.

MITCHELL, Tom M. **Machine Learning**. 1. ed. [*S.l.*]: McGraw-Hill Education, 1997. ISBN: 978-00-7042-807-2. Disponível em: https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf Acesso em: 28 abr. 2023.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003. Disponível em: https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf Acesso em: 12 abr. 2023

MORAN, J. **A educação que desejamos**: novos desafios e como chegar lá. Campinas: Papirus, 2007.

MUÑIZ, Luis Rodriguez *et al*. Dropout and transfer paths: what are the risk profiles when analyzing university persistence with *Machine Learning* techniques? **Plos One**, v. 14, n. 6, 2019. DOI: 10.1371/journal.pone.0218796.

NERI, M. C. **Motivos da evasão escolar**. 2009. Disponível em: https://www.cps.fgv.br/ibrecps/rede/finais/Etapa3-Pesq_MotivacoesEscolares_sumario_principal_anexo-Andre_FIM.pdf Acesso em: 25 fev. 2023.

OCDE. **Organização para a Cooperação e o Desenvolvimento Econômico**. Relatórios Econômicos OCDE: Brasil. OECD, fev. 2018. Disponível em: https://www.oecd.org/eco/surveys/Brazil-2018-OECD-economic-survey-overviewPortuguese.pdf. Acesso em: 25 fev. 2023.

PRESTES, Emília Maria da Trindade; FIALHO, Marília Gabriela Duarte. Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. **Ensaio: Avaliação de Políticas Públicas na Educação**, v. 26, n. 100, p. 869-889, 2018.

PRIMÃO, Aline Pacheco. **Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior:** um estudo no Instituto Federal de Santa Catarina. 2022. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/238320/PPAU0264-D.pdf?sequence=-1&isAllowed=y Acesso em: 25 fev. 2023.

RUSSEL, S.J.; NORVING, P. **Artificial Intelligence**: a modern approach. [*S.l.*]: Prentice-Hall, Inc., 1995.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. [*S.l.*]: Elsevier Campus, 2013.

SCHWARTZMAN, S. **A sociedade do conhecimento e a educação tecnológica.** Rio de Janeiro: IETS Instituto de Estudos do Trabalho e Sociedade. Universidade Federal do Rio de Janeiro - URFJ, 2005.

SILVA, Andréa Ferreira da; ALMEIDA. Aléssio Tony Cavalcanti de; RAMALHO, Hilton Martins de Brito. Predição do risco de reprovação no ensino superior usando algoritmos de *Machine Learning*. **Teoria e Prática em Administração**, v. 10, n. 2, p. 58-80, jul.-dez. 2020. DOI. 10.21714/2238-104X2020v10i2-51124.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. Florianópolis: Universidade Federal de Santa Catarina, 2005.

SILVA FILHO, R. L. L. *et al.* A evasão no ensino superior brasileiro. **Cadernos de Pesquisa,** São Paulo, v. 37, n. 132 p. 641-659, set./dez. 2005.

SILVA, F. C.; CABRAL, T. L. O.; PACHECO, A. S. V. 2020. **Evasão ou Permanência? Modelos Preditivos para a Gestão do Ensino Superior**. Disponível em: https://www.researchgate.net/publication/344884761_ Evasao_ou_Permanencia_Modelos_Preditivos_para_a_Gestao_do_Ensino_Superio r. Acesso em: 1 jun. 2023.

SILVA, Fernanda Cristina da; CABRAL, Thiago Luiz de Oliveira; PACHECO, Andressa Sasaki Vasques. Evasão ou permanência? Modelos preditivos para a gestão do ensino superior. **AAPE - Arquivos Analíticos de Políticas Educativas**, v. 28, n. 149, 2020.

SILVA, Raimunda Magalhães *et al*. **Estudos qualitativos:** enfoques teóricos e técnicas de coleta de informações. 2018. Disponível em: https://portais.univasf.edu.br/medicina-pa/pesquisa/ producao-cientifica/experiencias-qualitativas-ebook Acesso em: 1 jun. 2023.

WU, X. *et al*. Top 10 algorithms in data mining, knowledge information systems. **Know Inf Syst, Springer-Verlag,** v. 1, n. 14, p. 1-37, 2008.

YIN, Robert K. **Estudo de caso**: planejamento e métodos. Tradução de Daniel Grassi. 3. ed. Porto Alegre: Bookman, 2019.

ZHANG, J. *et al.* Early identification of dropout-prone students using *Machine Learning* models. **Educational Technology Research and Development,** v.67, p.225-240, 2019.

APÊNDICE A - QUESTIONÁRIO DE PESQUISA

A seguir apresenta-se o formulário eletrônico aplicado como instrumento de coleta de dados na pesquisa de campo:

Seção 1 de 2

Avaliação de variáveis que influenciam na evasão de alunos em instituição de ensino superior

Prezado e prezada, agradecemos por concordar em participar da nossa pesquisa.

O objetivo deste estudo é avaliar a importância das variáveis na evasão de alunos em instituições de nivel superior.

Este questionário é voluntário, tem 21 perguntas e leva cerca de 10 minutos para ser respondido. No início da pesquisa, será solicitado o seu endereço de e-mail. Garantimos que esta pesquisa segue um alto padrão acadêmico e os seus dados serão mantidos anonimamente. Não associaremos o seu nome (ou o nome da sua empresa) e endereço de e-mail às suas respostas e os usaremos exclusivamente para entrarmos em contato com você nas próximas etapas, caso seja necessário.

Este questionário é voluntário, tem 21 perguntas e leva cerca de 10 minutos para ser respondido. No início da pesquisa, será solicitado o seu endereço de e-mail. Garantimos que esta pesquisa segue um alto padrão acadêmico e os seus dados serão mantidos anonimamente. Não associaremos o seu nome (ou o nome da sua empresa) e endereço de e-mail às suas respostas e os usaremos exclusivamente para entrarmos em contato com você nas próximas etapas, caso seja necessário.

Ao clicar no botão "Próxima", você concorda em participar desta pesquisa. O consentimento também indica que você concorda que todas as informações coletadas individualmente poderão ser utilizadas por pesquisadores, de forma que a sua identidade pessoal seja preservada e protegida. Tal uso incluirá apresentações em reuniões científicas ou profissionais, publicação em revistas científicas, compartilhamento de informações anônimas com outros pesquisadores para verificação de precisão dos resultados do estudo e para futuras pesquisas aprovadas que tenham o potencial para aprimorar o conhecimento.

Para maiores informações não hesite em nos contatar através dos e-mails abaixo:

- Celso Barreto da Silva (cmmsoft@gmail.com): Mestrando em Sistemas e Computação pela Universidade Salvador - UNIFACS;
- Dra. Ana Patrícia Magalhães (anapatriciamagalhaes@gmail.com): Doutora em Ciências da Computação pela Universidade Federal da Bahia.

Este questionário estará disponível para respostas até o dia 26/05/2023.

Agradecemos pela sua atenção, participação e apoio.

E-mail *

E-mail válido

Este formulário está coletando e-mails. Alterar configurações

Você trabalha em uma instituição pública ou privada? *

- ○ Pública
- ○ Privada

Qual área de seu curso *

- ○ Exatas
- ○ Humanas
- ○ Saúde
- ○ Outros

Caso tenha marcado a opção outros, escreva neste espaço qual a sua área.

Texto de resposta longa

Seção 2 de 2

Por gentileza, responda as perguntas abaixo:

A seguir é apresentada uma lista das variáveis que podem influenciar na evasão de alunos no curso de nivel superior.

Atribua uma nota de 1 a 5 indicando a sua percepção sobre o grau de importância de cada uma dessas variáveis sobre a evasão de alunos.

Considere a seguinte escala para atribuir a sua nota:

1. nenhuma influencia na evasão,
2. pouca influencia na evasão,
3. influencia média na evasão,
4. influencia moderada na evasão,
5. bastante influencia na evasão.

Localização da Escola *

1	2	3	4	5
○	○	○	○	○

Gênero *

1	2	3	4	5
○	○	○	○	○

Idade *

1	2	3	4	5
○	○	○	○	○

Estado Civil *

1	2	3	4	5
○	○	○	○	○

Se o Estudante Trabalha? *

1	2	3	4	5
○	○	○	○	○

Educação dos Pais *

1	2	3	4	5
○	○	○	○	○

Renda dos Pais *

1	2	3	4	5
○	○	○	○	○

Nota média do Estudante *

1	2	3	4	5
○	○	○	○	○

Pontuação no Teste Final *

1	2	3	4	5
○	○	○	○	○

Pontuação no Teste Intermediário *

1	2	3	4	5
○	○	○	○	○

Frequência de Atendimento as Aulas *

1 2 3 4 5

Forma de Ingresso *

1 2 3 4 5

Tempo em Ingresso *

1 2 3 4 5

Turno *

1 2 3 4 5

Número de pessoas na família *

1 2 3 4 5

Em sua percepção, existe mais alguma outra variável que deveria ser abordada neste formulário? Qual? *

Texto de resposta longa

Qual o grau de importância desta variável sugerida? *

1	2	3	4	5
○	○	○	○	○

APÊNDICE B - VARIÁVEIS POR AUTOR

Segue a descrição das variáveis por autor utilizadas no contexto da pesquisa:

Autor	Variáveis
Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education (SUHARJITO, 2019)	Localização da Escola Gênero Idade Estado Civil O Estudante Trabalha? Educação dos Pais Renda dos Pais Nota média do Estudante Pontuação no Dever de Casa Pontuação no Teste Final Pontuação no Teste Intermediário Frequência de Atendimento as Aulas Número total de Créditos
Predição de Evasão Escolar na Licenciatura em Computação (de Jesus, H. O., Rodriguez, L. C., & Costa Junior, A. de O. (2021))	Período/Semestre Quanto Tempo de curso possui o aluno? N° de Disciplinas Aprovadas N° de Disciplinas Reprovadas N° de Disciplinas Canceladas N° de Disciplinas Trancadas N° Total de Disciplinas cursadas até então?
PREDIÇÃO DA EVASÃO ESCOLAR NOS CURSOS SUPERIORES DO IFMG -CAMPUS BAMBUÍ COM O APOIO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA (MELO, SOUZA, SANTOS, 2022)	Sexo Raça Forma de Ingresso Tempo em Ingresso Modalidade do Curso Classificação do Curso na CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Turno Carga Horária Quantidade de Aprovações e Reprovações Percentual de Reprovação no 1° Semestre O estudante recebe auxílio?
Pode a inteligência artificial apoiar ações contra evasão escolar universitária? (BITENCOURT, SILVA, XAVIER, 2022)	Nome do curso Idade se o estudante é cotista Forma de Ingresso Sexo Raça se o estudante trabalha em que Tipo de instituição o estudante completou o Ensino Médio Escolaridade dos Pais se o estudante Possui residência própria se o estudante mora em Área urbana ou rural Renda Familiar Número de pessoas na família Períodos Cursados Percentual de Aproveitamento de Créditos Coeficiente de Rendimento por Período Dependências Acumuladas

	Frequência Relativa no Período Nota Média no Período
Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil (Teodoro, L. A. & Kappel, M. A. A. (2020))	Foram usadas mais de 160 variáveis. Somente as 30 mais relevantes foram citadas. São elas, em ordem: Idade Número Total de Concluintes do Curso Se o aluno participa de Atividades Extracurriculares Carga Horária Total Raça Número Total de Matrículas do Curso Carga Horária Mínima Número Total de Ingressantes Número Total de Vagas Semestre de Referência Estado de Nascimento Se o aluno entrou pelo ENEM Tipo de Escola onde o Ensino Médio foi feito Nome do Curso Sexo Se o aluno recebe suporte financeiro Se o aluno entrou pelo vestibular Área do Curso Local da Instituição Se o aluno está na reserva de vagas Turno Quantidade de Computadores para alunos Se o aluno participa de pesquisas científicas Valor de despesa com pesquisas Número de professores do sexo masculino Número de funcionários técnico-administrativos Percentual de professores com graduação
Evasão ou Permanência? Modelos Preditivos para a Gestão do Ensino Superior (Silva, F. C., Cabral, T. L. O., & Pacheco, A. S. V. (2020))	Sexo Idade Cor Estado Civil Cidade de residência Cidade do Polo de Apoio Presencial Forma de Ingresso Pontos no Vestibular Índice de Aproveitamento Acumulado Tamanho da família Tempo de Deslocamento até o Polo em que Tipo de instituição estudou a maior parte do Ensino Médio Já fez Ensino superior antes? Já fez EAD antes? Frequência de uso do Computador Local de acesso à internet Tipo de conexão à internet Nível de Conhecimento Informático
Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? (MUNIZ, 2019)	Quantidade Pessoas na Família Horas de Estudo/Trabalho por Semana Idade Quantidade de Créditos Disputados Quantidade de Créditos Atendidos Quantidade de Créditos Passados Local de Residência Motivo dos Estudos Frequência de Atendimento as Aulas

	Dificuldade Participação de Atividades Não-Curriculares Gênero Nacionalidade Possui deficiência? Escolaridade dos Pais Profissão dos Pais Forma de Ingresso
Técnicas de Mineração de Dados: Um Estudo de Caso da Evasão no Ensino Superior do Instituto Federal do Maranhão (Gonçalves, T. C., da Silva, J. C., and Cortes, O. A. C. (2018))	Média de Faltas no Semestre Média de Notas no Semestre Percentual de Presença Número de Disciplinas no Semestre Tempo de Curso Sexo Turno Forma de Ingresso Ano Esperado para a Conclusão de Graduação Em que tipo de escola completou o Ensino Médio Estado Civil dos Pais Se os Pais Faleceram Renda Familiar Em que ano completou o Ensino Médio Se possui necessidades físicas ou visuais Renda per Capita Se trabalha Se é superdotado Grau de Instrução Estado Civil Número de Filhos Período Letivo Atual Período Letivo que começou o curso
Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD (Ramos, J., Silva, J., Prado, L., Gomes, A., and Rodrigues, R. (2018))	Média Semanal da Quantidade de Acessos ao Ambiente Quantidade de Acessos por Turno Tempo Médio de uso da Plataforma Quantidade de Acessos por Semestre Quantidade de Acesso as Atividades Quantidade geral e postagens e envio de mensagens Quantidade geral de recursos e atividades disponíveis para o aluno
Predição de Risco de Evasão de Alunos Usando Métodos de Aprendizado de Máquina em Cursos Técnicos (Bitencourt, P. and Ferrero, C. (2019))	Gênero Bairro Cidade Etnia Período de Ingresso Nome do Curso Turno do Curso Percentual de Faltas
Mineração de dados e Evasão estudantil: Analisando o curso de nível superior do Ifes (Gonc ̧alves, O. L. and Beltrame, W. A. R. (2019))	Frequência de Presença Cidade Escola de Origem Raça Renda Familiar Idade Tipo de Escola de Origem Sexo Forma de Ingresso

SARAIVA, *ET.AL*. 2019.	Idade Forma de Ingresso Sexo Coeficiente Rendimento Geral Renda Familiar Total frequencia Etnia Total Aprovado Desc. Estado Civil Total Aprovado Parcialmente Desc. Grau Instrução Total aprovado c/dependênci Estado Civil Pais Total Reprovado Grau de Escolaridade dos Pais Total de Matérias Trancadas Turno do Curso
Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática (Saraiva, D., Pereira, S., Gallindo, E., Braga, R., and Oliveira, C. (2019))	É calouro? Já repetiu o curso? É cotista? Etnia Sexo Grau de Escolaridade Faixa Etária Vive na região? (Mesmo Município)
Aplicação de técnicas de aprendizado de máquina em contexto acadêmico com foco na identificação de alunos evadidos e não evadidos (Soares, L. C. C. P., Ronzani, R. A., de Carvalho, R. L., and da Silva, A. T. R. (2020)).	Idade Sexo Estado Civil Raça Estado Natal Naturalidade Cidade Ano de Formação do 2° Grau Semestre de Ingresso Forma de Ingresso Último Semestre Cursado Campus Nome do Curso Turno do Curso
Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica (Amorim, M., Barone, D., and Mansur, A. (2008))	O ano e semestre de ingresso do aluno Número de Disciplinas Cursadas O percentual de aprovação do aluno no semestre anterior O percentual de desconto que o aluno possuía no semestre anterior A quantidade de prestações em aberto que o aluno possuía Coeficiente de Rendimento Escolar (CR) do aluno A Quantidade de disciplinas do curso; O percentual do curso que o aluno já havia completado;
Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS (Costa, S. S. D., Cazella, S., and Rigo, S. J. (2014))	ÚNICA variável usada para o estudo: desempenho do aluno no curso.

Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão (Silva, F., Silva, J., Silva, R., and Fonseca, L. (2015))	Nota Média em Fóruns Quantidade de Posts em Fóruns Total de Fóruns que o Aluno Participou Total de Postagens em todos os Fóruns Média de Postagens por Fórum Desempenho do Aluno nos Fóruns
A predictive model for identifying students with dropout profiles in online courses (Santana, M. A., de Barros Costa, E., dos Santos Neto, B. F., Silva, I. C. L., and Rego, J. B. (2015))	Resultado da Primeira Prova Uso do Blog Uso dos Fóruns Quantidade de Acesso a Plataforma Quantidade arquivos Enviados e Recebidos Cidade Uso da Wiki Uso do Glossário Estado Civil Gênero Salário
Predição da Evasão em Cursos de Graduação em Instituições Públicas (Kantorski, G., Flores, E., Schmitt, J., Hoffmann, I., and Barbosa, F. (2016))	Sexo Idade Estado Civil É cotista? Formação no Ensino Médio Já fez curso superior antes? Motivo de Escolha do curso Como vai se manter durante o curso Faz Atendimento Psicossocial Possui Bolsa Especializada? Ano de Ingresso Tempo no curso Período Atual Número de Disciplinas Aprovadas Número de Disciplinas Reprovadas Número de Disciplinas Reprovadas por Frequência Número de Disciplinas Dispensadas Número de Trancamentos Totais Número de Trancamentos Parciais Média de Notas Total de Disciplinas Matriculadas Média de Disciplinas por Período Moradia Estudantil Monitoria Media de utilização de instalação
Rede Bayesiana para Previsão de Evasão Escolar (Maria, W., Damiani, J., and Pereira, M. (2016))	Faixa Etária Nome do Curso Lotação da Turma Média de Nota Cidade Frequência nas Aulas Sexo Raça Possui Bolsa? Forma de Ingresso Possui Emprego?
predição de estudantes com risco de evasão em cursos técnico a distância (Queiroga, E., Cechinel, C., and Araujo, R. (2017))	Contagem de Interações Diárias na Plataforma Contagem de Interações Semanais Média, Mediana Desvio-Padrão das Interações

Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária (Paz, F. and Cazella, S. (2017))	Campus Incentivo (se a Faculdade Auxilia nas Contas) Data de Nascimento Semestre Atual Município onde Reside
Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados (Lanes, M. and Alcantara, C. (2018))	Faixa de Idade Média do ENEM Intervalo de Tempo desde Ensino Médio Último Coeficiente de Rendimento Área do curso Gênero Tipo de Escola de Origem É bolsista? Estado de Origem
Predicting Students Drop Out: A Case Study. (Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009))	Ano de Compleção do VWO O Currículo do VWO Número de Cursos Tomados no VWO Nota Média do VWO Nota Média em Matérias de Ciência no VWO Nota Média em Matérias de Matemática no VWO Número de Cursos de Matemática ou Ciência no VWO Tipo de Educação no HO Ano de Conclusão do HO Nota no HO Anos passados até entrada na faculdade
Mining Educational Data to Reduce Dropout Rates of Engineering Students (Pal, S. (2012))	Área do Curso Sexo Casta dos Estudantes Notas dos Estudantes no Ensino Médio e Senior Secondary Formato de Admissão Linguagem Usada em Estudo (Hindi ou Inglês) Local onde o estudante mora Renda Familiar Qualificação dos Pais Ocupação dos Pais
Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data (Marquez-Vera, C., Morales, C. R., and Soto, S. V. (2013))	Notas em Matemáticas, Inglês, Física, Humanas, Computação e Leitura Nota Média do Primeiro Exame Idade Nível de Motivação Deficiência Física Fuma?
Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program (Yukselturk, E., Ozekes, S., and Turel, Y. K. (2014))	Gênero Idade Nível Educacional Se teve experiência online anteriormente Ocupação Auto Eficácia (Self Efficacy) (?) Readiness (Preparação) (?) Conhecimento anterior Local de Controle (?)

Prediction of University Desertion through Hybridization of Classification Algorithms (Rocha, C. F., Zelaya, Y. F., Sanchez, D. M., and Pérez, F. A. F. (2017))	Ano de Entrada Gênero Último ano Escolar Vive com parentes? Pais separados? Tipo de Escola Anterior Se a casa é alugada Material da casa Possui computador? Forma de Ingresso Último ano de Curso Trabalha? Idade Renda Familiar Renda Pessoal Nota Mérito Acadêmico Peso da Nota Número de Cursos Aprovados Número de Cursos que Entrou Número de Cursos Reprovados
Modeling Students' Dropout in Mexican Universities (Rodrıguez-Maya, N. E., Lara-Alvarez, C., May-Tzuc, O., and Su´arez-Carranza, B. A. (2017)).	Média de Qualificação no Ensino Médio Percentil que o aluno ficou no Exame de Seleção (Vestibular) Quantos porcento acima do CNE (?) no Exame de Seleção Percentual de qualificação nas questões de raciocínio lógico matemático Posição do estudante no exame Qualificação nas questões de raciocínio verbal Qualificação nas questões de raciocínio verbal (CENEVAL index) Qualificação nas questões de raciocínio lógico matemático (CENEVAL index) CENEVAL index de qualificação no exame de seleção Qualificação em Matemática (CENEVAL index) Percentual de qualificação em Espanhol Qualificação de Matemática (CENEVAL index) Quantos estados visitou (no México) O pai fala algum dialeto indiano? Possui internet em casa? Qualificação nas questões de tecnologia da informação Possui TV a cabo? Qualificação nas questões de tecnologia da informação (CENEVAL index) Número de Computadores e Televisões em casa

Dropout Detection Using Non-Academic Data (Dharmawan, T., Ginardi, H., and Munif, A. (2018))	Gênero Distância da Casa Mora com Família Estado Civil Quantos membros na família? Tempo de Esperar para Estudo Possui internet em casa? Saúde Intensidade de uso do Celular Interesse em Pós-Graduação, Especialização Interesse na Graduação Motivação para Estudos Trabalha? Escolaridade dos Pais Ocupação dos Pais Relacionamento com a Família, Estudantes ou Docentes Personalidade
Perspectives to Predict Dropout in University Students with Machine Learning (Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M. (2018))	Gênero Mora na Universidade? Recebe Auxílio Recebe Empréstimo Tipo do Curso sendo Cursado Localização (Campus) Nome do Curso Turno do Curso Primeira Escolha de Carreira? Pediu mudança de carreira? Nota média do semestre Semestre Cursos que entrou no Semestre Cursos Aprovados no Semestre Cursos necessários para Graduação Semestres que não foram entrados anteriormente Ano de Entrada
Plataforma de Aprendizado de Maquina para Detecção e Monitoramento de Alunos com Risco de Evasão (Beltran, C., Xavier-Junior, J., Barreto, C., and Neto, C. O. (2019))	Sexo Idade Estado Civil Escola Modalidade do Curso Período de Ingresso Nota do Exame de Admissão (Vestibular) Total de Níveis no Curso Total de Créditos no Curso Total de Níveis Cursados Total de Créditos Cursados Período da Última Matrícula Último Nível Estudado Porcentagem CC Porcentagem NC Promedio do Último Período

	Semester and Group, Shift, Motivation Level, Administrative Sanction, No. of Friends, Additional Study Time, Study Form, Place of Study, When Studying, Doubts, Marital Status, Children, Religion, Chosen College Career, Influence on College Decision, Personality, Physical Disability, Serious Illness, Alcoholic Beverages, Smoking, Economic Status, Study Resources, Scholarship, Work, Who Lives with, Mother's Education Level, Father's Education Level, No. of siblings, Order of birth, Space to study, Stimulation of parents, Community inhabitants, Years living in community, Type of transportation, Distance to school, Attendance to classes, Bored in class, Considers useful knowledge, Difficult subject, Taking notes, Excess of homework, No. of students in group, Way of teaching, School infrastructure, Advisor, Interest of the institution.
Data Mining Applied in School Dropout Prediction Data Mining Applied in School Dropout Prediction (Viloria, A., Guliany, J. G., Nu´nez, W. N., Palma, H. H., and N˜ u´nez, L. N. (2020))	Age, Sex, Department of origin, School of origin regime, Secondary school model, Secondary school average, Mother's job, Father's job, No. of PC in family, Limited for exercise, Frequency of exercise, Time of exercise sessions, Grades in Logical Mathematical Reasoning, Grades in Mathematics, Grades in Verbal Reasoning, Grades in Spanish, Grades in Biology, Grades in Physics, Grades in Chemistry, Grades in History, Grades in Geography, Grades in Civic Formation, Grades in Ethics, Grades in English and Grade of EXANI I. Grade in Humanities, Grade in Reading and Writing Workshop, Grade in English, Grade in Computer, Academic Status, Grade in Mathematics, Grade in Physics, Grade in Social Sciences.