Mineração de Dados Educacionais: Uma Análise Sobre os Preditores de Evasão no Ensino Superior*

Felipe Silvestri[1], Andrws Vieira[1], Vanessa Faria de Souza[2]

[1] Instituto Federal de Ciência e Tecnologia do Rio Grande do Sul - Campus Ibirubá
Rua Nelsi Ribas Fritsch, 1111 – CEP: 98200-000 – Ibirubá – RS – Brasil
[2] Instituto Federal de Ciência e Tecnologia do Rio Grande do Sul - Campus Vacaria
Rua Eng. João Viterbo de Oliveira – CEP: 95200-000 – Vacaria – RS – Brasil

Resumo. *A evasão no ensino superior é um desafio significativo para instituições de ensino, impactando tanto os estudantes quanto a sociedade. Este artigo investiga os principais preditores de evasão acadêmica por meio da aplicação de técnicas de Mineração de Dados Educacionais (MDE). A pesquisa utiliza um conjunto de dados provenientes de uma instituição de ensino superior, explorando características como desempenho acadêmico, perfil socioeconômico, engajamento estudantil, dados sobre a família, dentre outros. Foi utilizada como técnica de Mineração de Dados algoritmos de Aprendizagem de Máquina. Estes, além de possibilitarem a construção de um modelo eficaz para predição da evasão dos alunos, possibilitaram a identificação dos fatores mais relevantes para prever a evasão. Os resultados destacam que variáveis relacionadas ao rendimento acadêmico são os principais indicadores de risco. As conclusões oferecem insights que podem orientar estratégias institucionais para mitigar a evasão, promovendo a permanência e o sucesso acadêmico. Palavras-chave: Mineração de Dados Educacionais; Aprendizagem de Máquina; Evasão no Ensino Superior.*

Abstract. *Dropout rates in higher education pose a significant challenge for educational institutions, affecting both students and society at large. This article explores the key predictors of academic dropout through the application of Educational Data Mining (EDM) techniques. Utilizing a dataset from a higher education institution, the research examines various factors, including academic performance, socioeconomic status, student engagement, and family background. Machine Learning algorithms were used as part of the Data Mining techniques, allowing for the development of an effective model to predict student dropout rates while also identifying the most relevant contributing factors. The findings emphasize that variables related to academic performance serve as the primary risk indicators. The conclusions provide valuable insights that can inform institutional strategies aimed at reducing dropout rates and fostering student retention and academic success. Keywords: Educational Data Mining; Machine Learning; Dropout in Higher Education.*

*Trabalho de Conclusão de Curso (TCC) do curso Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul – Campus Ibirubá, 2025

1. Introdução

Nas duas últimas décadas, a educação superior brasileira foi marcada por forte expansão sob todos os aspectos, como por exemplo, o aumento do número de instituições, de cursos, de vagas, de ingressantes, de matrículas e de concluintes (RISTOFF, 2014). Com o crescimento das universidades brasileiras por meio de programas como o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI) e, também, do Programa Universidade para Todos (PROUNI), houve a implementação de novas políticas públicas no ensino superior (ALVES; GAYDEZKA; CAMPOS, 2018). Para tanto, foram necessários grandes investimentos em infraestrutura como a criação de novos campi em todo o Brasil, bem como a oferta de novos cursos e consequentemente, o aumento da oferta de vagas no ensino superior público (ALVES; GAYDEZKA; CAMPOS, 2018).

Quando se fala em expansão do ensino superior, seja ele público ou privado, se faz necessário discutir sobre evasão. Para Fritsch, Rocha e Vitelli (2015) é considerado evasão no Ensino Superior, o ingresso e a não conclusão de um curso de graduação por desistência. Para os autores, trata-se de um processo de exclusão determinado por fatores e variáveis internas e externas às Instituições de Ensino Superior (IES). A evasão é, certamente, um dos problemas que mais afligem as instituições de ensino em geral (FILHO et al., 2007). A compreensão de suas causas tem sido objeto de muitos trabalhos e pesquisas educacionais. Para Lobo (2012), no ensino superior, a evasão estudantil:

> [...] é um problema internacional que afeta o resultado dos sistemas educacionais. As perdas de estudantes que iniciam, mas não terminam seus cursos são desperdícios sociais, acadêmicos e econômicos. Para o setor público, são recursos públicos investidos sem o devido retorno, já para o setor privado, é uma importante perda de receitas. Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico (LOBO, 2012, p. 642).

Para Mello et al. (2013), a evasão pode ser dividida em fatores internos e externos. Os fatores internos geralmente estão ligados à universidade, dentre eles podemos citar a desistência do curso pelo descontentamento com os métodos didáticos pedagógicos ou com a infraestrutura da universidade. Já os externos, são aqueles vinculados ao próprio discente, como a dificuldade de adaptação ao ambiente universitário, problemas de ordem financeira e pessoal ou o curso escolhido não era o que o discente esperava. Os elementos relacionados à evasão e seus motivos, é algo que merece atenção e devem ser objeto de estudo e de preocupação das instituições de ensino superior, principalmente, no que tange a coleta de dados no momento em que o discente realiza seu pedido de desligamento da universidade (ALVES; GAYDEZKA; CAMPOS, 2018).

A alta taxa de evasão no ensino superior representa um desafio significativo, de acordo com o Instituto Semesp (2023)[1], criado pela Secretaria de Modalidades Especiali-

[1] O Instituto Semesp é um centro de inteligência analítica criado pelo Semesp. Integrado por especialistas com sólida experiência no levantamento e análise de dados sobre o ensino superior, o Instituto desenvolve estudos, pesquisas, indicadores e análises estatísticas referentes ao setor. Seu objetivo é disponibilizar para pesquisadores, educadores, gestores privados e públicos, jornalistas e para a sociedade em geral informações relevantes e confiáveis que lhes permitam tomar decisões, estabelecer estratégias ou formular políticas públicas, visando o desenvolvimento da educação superior.

zadas de Educação, cerca de 21% dos alunos evadiram o ensino superior em instituições Públicas no ano de 2020 e 20% em 2021, em se tratando de instituições particulares foi cerca de 30% no biênio, e para a modalidade de Ensino a Distância (EaD) os índices para os dois anos ficaram em torno de 33% e 36% na rede privada e 31% e 27% na rede pública, como pode-se ser visualizado na Figura 1.

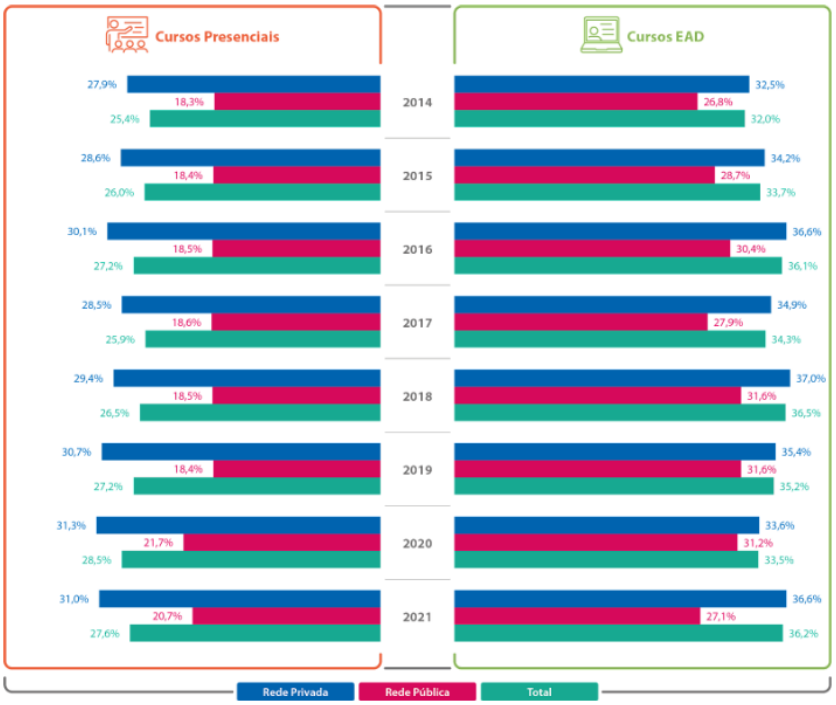

Figura 1. Evasão no Ensino Superior. Fonte: Semesp (2023)

Dessa forma, compreender os fatores principais que levam os alunos a abandonarem os cursos torna-se crucial. Além disso, é importante ressaltar que a evasão no ensino superior não apenas impacta os indivíduos diretamente envolvidos, mas também tem repercussões sociais e econômicas. A pesquisa e compreensão desse fenômeno, ainda mais depois da pandemia, em que percebe-se que a evasão se acentuou, tornou-se uma tarefa importante, sobretudo para promover políticas e práticas educacionais mais eficazes.

Neste contexto, este projeto tem como propósito abordar essa problemática, visando analisar dados de alunos, que contêm informações a respeito de aspectos internos e externos durante a realização de um curso de ensino superior, estes dados estão organizados em forma de variáveis contidas em uma base de dados, previamente coletada. O objetivo é determinar quais desses elementos exercem maior influência sobre a evasão dos estudantes. Para isso, pretende-se utilizar a mineração de dados educacionais, a partir de uma das principais técnicas para análise de dados – a Aprendizagem de Máquina (AM) – aplicando algoritmos de aprendizagem supervisionada, isso porque a base utilizada está rotulada, então é possível saber quais alunos evadiram, ou não, de seus respectivos cursos. Dessa forma, será possível formatar uma base de dados, apenas com as informações mais relevantes, no que tange a evasão dos alunos, e com a aplicação dos algoritmos de AM sobre esta base, deve ser possível gerar um modelo que acredita-se ser capaz de predizer a evasão de alunos no ensino superior, de maneira eficaz.

A partir dessa delimitação do tema, é possível estabelecer a questão de pesquisa que irá nortear este estudo: Quais são as variáveis que mais impactam na evasão de um aluno no ensino superior? Para responder a essa questão, como já destacado, será utilizado como alicerce a mineração de dados educacionais e suas técnicas para análise de dados, sendo possível estabelecer um modelo de dados otimizado para a predição da evasão em cursos de ensino superior. Com as predições realizadas pelo modelo, acredita-se obter indicações de alunos com pretensão à evasão; tais informações são úteis para gestores de instituições de ensino, que podem fazer intervenções antecipadas, prevenindo a evasão de curso.

O objetivo geral deste estudo é identificar quais são os principais atributos que implicam na evasão de alunos do ensino superior, por meio de técnicas da mineração de dados educacionais, enfocando a aprendizagem de máquina. Para alcançar o objetivo específico destacado, é necessário atingir algumas metas, que são caracterizadas como os objetivos específicos deste estudo:

- Analisar uma base de dados de alunos do Ensino Superior;
- Aplicar e avaliar algoritmos de Aprendizagem de Máquina (Naive Baïes, Árvore de Decisão, *Random Forest*, *Support Vector Machines* (SVM), Regressão Logística e Redes Neurais) na perspectiva de selecionar o mais eficaz;
- Realizar testes sobre a base de dados, subdividindo a mesma, para verificar qual o conjunto de atributos que mais impactam na evasão do aluno;
- Selecionar e descrever os atributos identificados como os mais significativos na evasão dos alunos;
- Gerar um modelo eficaz para a predição da evasão de alunos no ensino superior.

Diante do exposto este artigo está organizado da seguinte forma: nesta seção foi tratada da contextualização e objetivos do estudo; na seção seguinte serão delimitadas conceituações relevantes para o desenvolvimento desta pesquisa; na seção três alguns trabalhos relacionados são detalhados; na sequência é apresentada a metodologia utilizada; a seção cinco apresenta os resultados alcançados; a seção seis traz as discussões; e por fim a seção sete apresenta as considerações finais deste artigo.

2. Revisão Bibliográfica

Nesta seção serão elucidados conceitos importantes para o desenvolvimento e compreensão deste estudo, dentre eles: a Mineração de Dados Educacionais, técnicas para mineração de dados – Aprendizagem de Máquina, algoritmos para as análises de dados e, avaliação de algoritmos.

2.1. Mineração de Dados Educacionais

A Mineração de Dados é definida como o estudo de coleta, limpeza, processamento, análise e obtenção de informações e ideias úteis de dados. Há, entretanto, uma grande diversidade em termos de domínios problemáticos, aplicativos, formulações e representações de dados encontradas em aplicativos reais. Então, "Mineração de dados" é um termo abrangente para descrever diferentes aspectos de processamento de dados (AGGARWAL, 2015). A utilização das técnicas de mineração de dados para extrair informações pertinentes de conjuntos diversificados de dados educacionais, é conceituado

como Mineração de Dados Educacionais (MDE), que segundo a Sociedade Internacional de Mineração de Dados Educacionais[2], pode ser definida da seguinte forma:

> É uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar dados únicos e cada vez mais em larga escala, provenientes de contextos educacionais e usa esses métodos para entender melhor os alunos e as configurações em que aprendem (EDM, 2020).

Reforçando as definições acima, Bakhshinategh et al. (2018) salienta que a mineração de dados educacionais é o campo de uso das técnicas de mineração de dados em ambientes educacionais. Segundo Souza (2021) a Mineração de Dados educacionais tem as seguintes etapas (Figura 2):

1. *Definição da função da MDE* – Nessa etapa é realizada a determinação do objetivo do processo MDE, para qual finalidade ela está sendo aplicada, como por exemplo: identificação de padrões, detecção de desvio, segmentação, sistemas de recomendação, análise de ligações e regras de associação, sumarização e visualização, mineração de textos, afinidade em grupos, descrição de grupos, para isso é preciso especificar que tipo de conhecimento pretende-se extrair dos dados.

2. *Formatação dos dados que serão utilizados* – Diversos tipos de armazenamentos de dados e de bancos de dados podem ser manipulados no processo de mineração, cabe ao usuário definir qual formato é o mais adequado para aplicação das técnicas de mineração selecionadas, outro ponto importante é que baseado no tipo de conjunto de dados disponível para análise é que definem-se os padrões, relacionamentos ou informações que se consegue minerar. Nessa etapa todas as incoerências na base devem ser corrigidas e caso for necessário podem ser acrescentados mais atributos que sejam resultantes da combinação de outros, ou que possam ser deduzidos de outros, como a idade que pode ser calculada a partir da data de nascimento, ou o total de atividades realizadas que pode ser efetuada por meio de uma soma.

3. *Definição das Técnicas de MDE* – A definição das técnicas é um passo importante, pois elas devem ser específicas para o tratamento da função de MDE estabelecida, as técnicas mais utilizadas para MDE são: a Estatística Descritiva, a Aprendizagem de Máquina e mais recentemente tem sido empregada a Aprendizagem Profunda. Cabe salientar que cada uma dessas técnicas possui algoritmos, rotinas e/ou procedimentos específicos para manipulação dos dados.

4. *Delineamento de como essas técnicas serão aplicadas* – Nessa etapa são selecionadas as ferramentas que vão dar suporte ao desenvolvimento de sistemas capazes de processar os dados e gerar os resultados esperados.

[2]https://educationaldatamining.org/

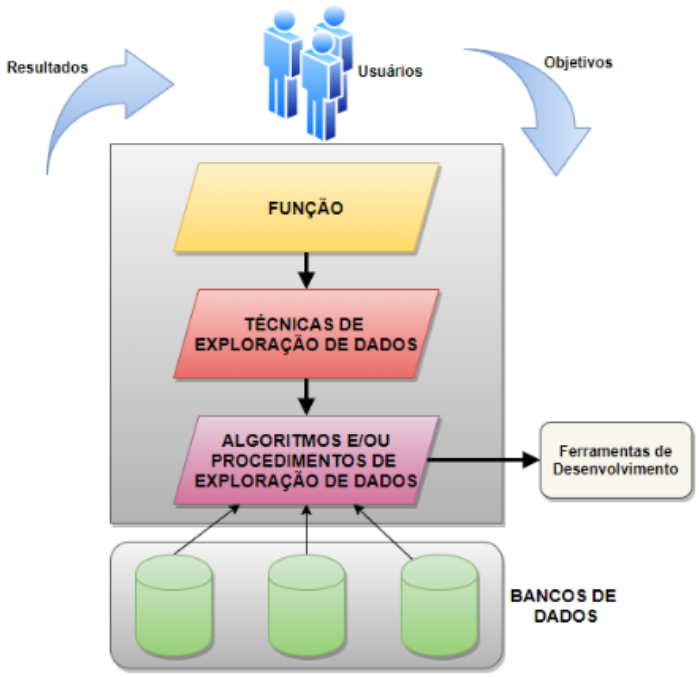

Figura 2. Etapas do processo de MDE. Fonte: Souza (2021)

Neste estudo, a primeira etapa (definição da função da MDE) corresponderá a definição do objetivo da pesquisa, neste sentido o intuito da MDE é fornecer um modelo para predição da evasão de alunos, que possa auxiliar no processo de permanência e êxito, amenizando a evasão por meio de intervenções precoces entre os alunos propensos a evadir. A segunda etapa, que corresponde a formatação dos dados que serão utilizados, será realizada por meio de adequações a base de dados que já existe para aplicação das técnicas para análise de dados. Quanto à terceira etapa, será utilizada a Aprendizagem de Máquina como técnica para a MDE. Por fim, essa técnica será aplicada por meio de algoritmos de aprendizagem supervisionada, utilizando a linguagem de programação R.

2.2. Técnicas para Mineração de Dados Educacionais

Algumas das principais técnicas para análise de dados são: Estatística Descritiva, a Aprendizagem de Máquina e mais recentemente tem sido empregada a Aprendizagem Profunda. Essas técnicas possuem algoritmos, rotinas e/ou procedimentos específicos para manipulação dos dados. Neste estudo será utilizada a Aprendizagem de Máquina como principal técnica para descobrir padrões nos dados, dessa forma esta será definida na sequência.

2.2.1. Aprendizagem de Máquina

O conceito de Aprendizagem de Máquina trata da extração de informações dos dados. É uma área de pesquisa formada pela interseção de Estatística, Inteligência Artificial e Ciência da Computação. Esta, muitas vezes é referenciada como análise preditiva ou aprendizado estatístico e muitos pesquisadores defendem que ela é um campo

da Inteligência Artificial (BISHOP, 2006),(HASTIE; TIBSHIRANI; FRIEDMAN, 2009), (MITCHELL, 1997 apud SOUZA, 2021).

Para Navarro et al. (2017) a técnica de AM oferece soluções para automatizar a análise de *Big Data*, os autores afirmam que ela pode ser considerada como um conjunto de métodos que podem detectar automaticamente padrões nos dados, então pode-se utilizar esses padrões descobertos para fazer previsões ou para tomada de decisão. Existem dois tipos principais de AM, a aprendizagem supervisionada e a não supervisionada.

Na Aprendizagem Supervisionada, a instância de entrada e a categoria correspondente à qual essas instâncias pertencem são fornecidos para o algoritmo. O algoritmo de AM assimila a relação entre a entrada e a saída e, em seguida, prevê a saída para amostras de dados de entradas cujas saídas não foram fornecidas (SOUZA, 2021).

A Aprendizagem Supervisionada é utilizada para resolver dois tipos diferentes de problemas: Classificação e Regressão, também chamadas de funções de Aprendizagem Supervisionada. A Classificação refere-se ao processo de previsão de valores de saída discretos para uma entrada, por exemplo, dado uma entrada o algoritmo de classificação prevê se um e-mail é *spam* ou legítimo, se um aluno será aprovado ou reprovado no exame. Enquanto na Regressão, a tarefa do modelo de aprendizado de máquina é prever um valor contínuo, por exemplo para certa entrada os algoritmos de regressão podem prever o preço de uma casa (SOUZA, 2021).

No que tange a Aprendizagem não Supervisionada, os algoritmos recebem conjuntos de dados desprovidos de rótulos. Nesse contexto, cabe ao algoritmo a tarefa de identificar padrões nos registros e agrupar elementos com características semelhantes. Geralmente, a maioria dos conjuntos de dados do mundo real não possui classificações pré-estabelecidas; assim, a abordagem não supervisionada pode servir como uma etapa inicial para preparar dados que posteriormente serão utilizados em métodos de aprendizado supervisionado (SOUZA, 2021).

A Aprendizagem Não Supervisionada engloba a Redução de Dimensionalidade e Agrupamento. A Redução de Dimensionalidade busca reduzir a dimensão de dados visando economizar recursos computacionais, mantendo a eficácia dos algoritmos. O Agrupamento envolve a categorização não supervisionada de objetos semelhantes, identificando novas categorias, mesmo quando a similaridade entre atributos não é evidente (SOUZA, 2021)

A Aprendizagem de Máquina (AM) utiliza algoritmos que aprendem padrões nos dados. Este estudo emprega algoritmos supervisionados de classificação, descritos a seguir.

2.2.2. Algoritmos de Aprendizagem de Máquina

Na aprendizagem de máquina supervisionada podem existir dois tipos de problemas, de regressão e de classificação. Na classificação temos categorias ou classes - caracterizando cada instância da base de dados; e na regressão temos valores numéricos como sendo os rótulos. A base de dados aqui analisada, está classificada com rótulos que indicam se o aluno evadiu, ainda cursa ou concluiu o curso, dessa forma se caracterizando como um problema de classificação. Para tanto, os algoritmos que serão utilizados neste

estudo são algoritmos supervisionados para classificação, pois seu conceito é sobre prever um rótulo e a regressão é sobre prever um valor continuo. Existem alguns algoritmos de AM que já estão consolidados na literatura, pretende-se utilizar estes como principal forma para analisar os dados educacionais, extrair as informações relevantes sobre a evasão dos alunos, os algoritmos que serão utilizados neste estudo são definidos na Tabela 1.

Tabela 1: Algoritmos e suas definições

ALGORITMO	**DESCRIÇÃO**
Naïve Bayes	O algoritmo Naïve Bayes é um algoritmo supervisionado de aprendizado de máquina baseado no teorema de Bayes e fundamentado no princípio de independência de recurso, que afirma que os recursos de um conjunto de dados não têm relação entre si. Este algoritmo não se preocupa se esses recursos dependem um do outro, a fruta é declarada como banana por contribuição independente desses recursos. Devido a essa suposição de independência, o algoritmo tem essa denominação de ingênuo e é o mais simples de todos os algoritmos de aprendizado de máquina e, no entanto, é muito aplicado por ser eficaz.
Árvore de Decisão	A Árvore de Decisão é um algoritmo de Aprendizado de Máquina fundamentado na entropia, opera sob o conceito em que cada atributo presente no conjunto de dados é considerado como um nó na árvore. A dinâmica do processo envolve a tomada de decisões em cada nó, determinada pelo valor do atributo específico naquele ponto da árvore. Esse procedimento é repetido até que o nó da folha seja alcançado, pois é nesse ponto que reside a decisão final que conduz à classificação da instância em questão.
Random Forest	Uma única Árvore de Decisão pode ser tendenciosa, dependendo dos dados. Para mitigar esse viés, uma abordagem eficaz é utilizar várias árvores de decisão, cada uma fazendo sua própria previsão. A previsão final é obtida calculando a média de todas as previsões das árvores, conhecida como "ensemble learning"(aprendizado em conjunto). Esse método combina algoritmos semelhantes ou diferentes para aumentar a capacidade do modelo de Aprendizado de Máquina. Um exemplo desse enfoque é a Floresta Aleatória, que une múltiplos algoritmos de Árvore de Decisão para formar uma floresta.

SVM	O Support Vector Machines destaca-se como uma ferramenta robusta na construção de classificadores. Visa estabelecer uma fronteira de decisão entre duas classes, permitindo a previsão de rótulos com base em um ou mais vetores de características. Funciona no intuito da regressão linear em um espaço de recurso bidimensional, é encontrar uma linha reta que separa com êxito os dados de diferentes classes, no entanto no mundo real, pode haver vários limites de decisão que podem classificar os dados com êxito.
Regressão Logística	A regressão logística é uma técnica de análise de dados que usa matemática para encontrar as relações entre dois fatores de dados. Em seguida, essa relação é usada para prever o valor de um desses fatores com base no outro. A previsão geralmente tem um número finito de resultados, como sim ou não.
Redes Neurais	São algoritmos que executam transformações matemáticas nas bases de dados que recebem para processamento. Em cada nó de cada camada, ocorre uma multiplicação entre o valor de entrada e o peso associado ao nó correspondente. Em seguida, soma-se o viés associado a esse nó, e o resultado é encaminhado adiante, seguindo a concepção de feed forward (avançar). O viés desempenha um papel semelhante ao intercepto adicionado em uma equação linear; age como um parâmetro adicional ajustando a saída em conjunto com a soma ponderada das entradas para o perceptron. Trata-se de uma constante que auxilia o modelo a adaptar-se melhor aos dados fornecidos.

2.2.3. Avaliação de Algoritmos de Aprendizagem de Máquina

Para realizar a verificação dos resultados de um modelo de classificação ou regressão são necessários dois itens: os métodos de avaliação e as métricas de interpretação, os dois devem ser aplicados em conjunto para que seja possível observar se um modelo é eficaz ou não. Os métodos indicam como esse modelo será avaliado, e as métricas traduzem os resultados da aplicação desses métodos em números que possam ser interpretados. Os dois principais métodos de avaliação são Treinamento e Teste, e Validação Cruzada.

No Treinamento e Teste a base de dados é dividida de forma aleatória em duas porções, uma para treinamento e outra para teste, de acordo com Japkowicz e Shah (2014) geralmente fica 85% das instâncias para treinamento e 15% para teste. O método de Validação Cruzada, citado por Japkowicz e Shah (2014), é uma abordagem detalhada e precisa. O usuário escolhe o número de partes para dividir os dados (geralmente 5 a 10). Em cada iteração, o método seleciona aleatoriamente uma parte para teste e as restantes para treinamento. Esse processo é repetido até que todas as partições tenham sido usadas

como conjunto de teste. A eficácia é calculada em cada iteração e, ao final, a média desses resultados fornece a eficácia geral do modelo.

A aplicação do método de avaliação por si só não determina a eficácia do modelo. É necessário empregar métricas que permitam interpretar a precisão das classificações. As métricas mais comuns para avaliação de modelos de Aprendizagem Supervisionada estão apresentadas na Tabela 2.

Tabela 2: Métricas de avaliação

MÉTRICA	**DESCRIÇÃO**
Precisão da Classificação (Acurácia)	A precisão é uma métrica de avaliação comum para problemas de classificação. É o número de previsões corretas feitas como uma proporção de todas as previsões realizadas sobre a base de testes. Em outras palavras, é a porcentagem de instâncias classificadas corretamente de todas as instâncias. Pode ser considerada mais útil em uma classificação binária do que em problemas de classificação de várias classes, porque pode ser menos claro exatamente como a precisão se divide nessas classes.
Intervalo de Confiança (IC)	Corresponde a uma métrica que indica que há uma probabilidade de 95% que a verdadeira precisão do modelo algorítmico testado esteja dentro desse intervalo.
Taxa de não informação	Essa é a precisão alcançável, sempre prevendo o rótulo da classe majoritária. Portanto, corresponde à melhor escolha, sem outras informações.
Valor de P	Consiste em um teste unilateral para verificar se a precisão é melhor que a taxa de não informação, considerando a maior porcentagem da classe dos dados.
Kappa	Corresponde a uma medida de concordância usada em escalas nominais que fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando assim o quão legítimas as interpretações são. É parecida com a precisão, excetuando por ser normalizada na linha de base do acaso no conjunto de dados. É passível de ser considerada uma medida mais utilizada para problemas com desequilíbrio nas classes.

Os métodos de avaliação citados nesta seção são utilizados tanto para problemas de Classificação como de Regressão, no que diz respeito às métricas são utilizadas em especial no contexto da Classificação, embora a acurácia e a matriz de confusão sejam utilizados para ambas as abordagens.

3. Trabalhos Relacionados

Com os avanços tecnológicos, os estudos na área da Mineração de Dados Educacionais (MDE) se tornam cada vez mais intensos, para alcançar, por exemplo, objetivos como o deste trabalho. Diversos trabalhos vêm sendo desenvolvidos, porém este assunto ainda tem grande potencial evolutivo. No decorrer das pesquisas realizadas para desenvolver este estudo, várias pesquisas correlatas foram identificadas, dentre estas destacam os estudos de Martins et al. (2021), Souza (2020) e Hegde (2018).

A investigação de Martins et al. (2021) desenvolveu modelos de classificação utilizando Aprendizagem de Máquina para prever e identificar previamente estudantes propensos a não concluírem sua graduação. Neste estudo foi utilizado uma base de dados proveniente do Instituto Politécnico de Portalegre, em Portugal. Foram aplicados durante o processo do trabalho algoritmos consolidados na literatura (Regressão Logística, *Support Vector Machine* (SVM), Árvore de Decisão e Floresta Aleatória) e também um modelo denominado *Boosting Model*, que aumenta a precisão geral do sistema. O trabalho alcançou seus objetivos, e os autores puderam constatar que os algoritmos de *Boosting* tiveram melhor resposta se comparado com os algoritmos padrões, mesmo tendo algumas falhas ao longo do processo.

O estudo desenvolvido por Souza (2020) teve como objetivo realizar a previsão do desempenho de alunos realizando um comparativo entre as técnicas consolidadas na área da MDE e a técnica de AP (Aprendizagem Profunda). Os dados utilizados na pesquisa foram retirados de um repositório público UCI *Machine Learning*, provenientes de relatórios escolares e questionários. Estes dados eram compostos por dois conjuntos de dados de 1044 alunos, contendo seus desempenhos nas disciplinas de Matemática e Português. O processo do trabalho se desenvolve em quatro fases: Na primeira fase foram coletados os dados do repositório público. Após a coleta, os dados foram sistematizados em um *Data Frame*. Na segunda fase foi realizado o pré-processamento e transformação dos dados. Na terceira fase foram aplicados os algoritmos de Aprendizagem de Máquina definidos. Na quarta fase foram analisados os resultados baseados nas métricas definidas pela autora do trabalho.

O trabalho alcançou seu objetivo de realizar a previsão do desempenho dos alunos, e identificar quais são os principais atributos que dão melhor suporte na previsão. As redes neurais tiveram a maior acurácia, alcançando 94% de acerto do total das previsões feitas e com base nas técnicas de MDE aplicadas, os atributos que mais influenciaram nas previsões foram os relacionados ao desempenho escolar dos alunos, ou seja, relacionam-se a fatores internos a Instituição de Ensino.

Por fim, destaca-se o estudo realizado por Hegde (2018) este teve como objetivo demonstrar o uso da Mineração de Dados Educacionais, por meio de modelos para prever a evasão de alunos. A base de dados adquirida é fruto de um questionário de pesquisa booleano (sim/não), composto por 54 perguntas. Após o pré-processamento dos dados, foi utilizado o algoritmo de aprendizagem de máquina Naive Bayes nesta base. Este trabalho cumpriu sua meta de desenvolver um modelo para prever, de forma precoce, a possível evasão de alunos de seus respectivos cursos. A aplicação do algoritmo Naive Bayes na base de dados alcançou uma acurácia de 72% do total de previsões realizadas.

Ao analisar os estudos correlatos foi possível identificar quais são as técnicas mais

empregadas para a realização do processo de MDE, sendo possível identificar que algoritmos como: Árvore de Decisão, SVM e Naïve Bayes, têm sido bastante utilizados para a realização de previsões em bases de dados educacionais. Além disso, foi possível identificar métricas importantes para análise de tais técnicas, como a acurácia e a matriz de confusão. Ademais, pode-se ter uma noção de quais são as melhores técnicas, ficando claro que no estudo de Martins et al. (2021) o modelo de *Boosting* foi o mais eficiente, em Souza (2020) foram as Redes Neurais Artificiais Multicamadas (RNAM) as mais precisas e em Hegde (2018), mesmo sendo aplicado apenas o algoritmo Naive Bayes, foi possível perceber um bom desempenho deste modelo. Uma melhor comparação entre os estudos analisados pode ser observada na Tabela 3.

Tabela 3: Estudos sobre Mineração de Dados Educacionais

Estudo	**Objetivo**	**Algoritmos Utilizados**	**Métricas de Avaliação**	**Base de Dados (Origem)**
MARTINS et al. (2021)	Construir modelos de classificação com Aprendizagem de Máquina para prever e identificar previamente estudantes propensos a não concluírem sua graduação.	Regressão Logística, SVM, Árvore de Decisão, Random Forest	Falso Negativo, Falso Positivo real, Falso Positivo, Média entre falso negativo e falso positivo, Acurácia	Instituto Politécnico de Portalegre, Portugal (Dados coletados entre 2008 e 2018).
SOUZA (2020)	Realizar a previsão do desempenho de alunos, com a intenção de disponibilizar um documento que apresenta de maneira detalhada como realizar o processo de MDE (Mineração de Dados Educacionais).	Naïve Bayes, Árvore de Decisão, SVM, RNAM	Matriz de Confusão, Acurácia, IC, Taxa de não informação, Valor de P, Kappa, AUC, Valores Preditivos Positivo e Negativo, Prevalência, Taxa de Detecção, Prevalência de Detecção, Precisão Balanceada	Escola de Ensino Médio de Portugal – Disciplinas de Matemática e Português.

HEGDE (2018)	Demonstrar o uso da Mineração de Dados Educacionais através dos modelos de implementação de aprendizagem de máquina para prever a evasão de alunos.	Naïve Bayes	Matriz de Confusão, Kappa, IC	Através de um questionário de pesquisa booleana contendo 54 questões, utilizando o Google Forms.

4. Metodologia

Esta seção tem como finalidade apresentar o processo que foi empregado para o desenvolvimento deste trabalho. O objetivo principal desta pesquisa foi identificar quais eram os principais atributos dentro de uma base de dados que influenciavam na evasão de alunos no Ensino Superior. Este objetivo caracterizou esta pesquisa como explicativa, pois teve como preocupação central identificar os fatores que determinaram, ou que contribuíram para a ocorrência de fenômenos (GIL, 2002). Seus procedimentos foram baseados nos conceitos e nos procedimentos da pesquisa ex-post facto, a partir dos fatos passados, dado que este estudo se deu após os acontecimentos que contribuíram para a construção da base de dados (GIL, 2002). Para tanto, o projeto foi dividido em sete etapas:

1. *Realização da Revisão Bibliográfica:* Para que a realização deste trabalho fosse feita de forma satisfatória, foi necessária a busca de conhecimento em livros, documentos e artigos, para que fosse criado um embasamento teórico sólido dos conceitos que foram abordados durante o desenvolvimento do tema escolhido.

2. *Realização do pré-processamento da base de dados:* Em sua grande maioria, quando os dados foram coletados, eles não se encontravam em um formato adequado para o processamento, inviabilizando a interpretação dos algoritmos de aprendizagem que foram aplicados nela. Nesta etapa, os dados contidos na base foram adequados, realizando-se, se necessário, a limpeza dos dados, que consistiu em corrigir dados faltantes ou incorretos, para que fosse possível aplicar os algoritmos de Aprendizagem de Máquina escolhidos para este trabalho. A base de dados analisada foi uma base de acesso público, proveniente do Instituto Politécnico de Portalegre, Portugal, com dados coletados entre 2008 e 2018, contendo 37 variáveis/atributos e 4.424 instâncias. A base de dados pode ser acessada no portal UCI Machine Learning[3].

3. *Análise descritiva dos dados:* Foram analisadas estatísticas descritivas a respeito das informações que puderam ser extraídas da base de dados.

4. *Aplicação dos algoritmos de Aprendizagem de Máquina e avaliação dos algoritmos:* Após a realização do pré-processamento dos dados, foi feita a aplicação de seis algoritmos de Aprendizagem de Máquina escolhidos para este estudo (Naive Bayes, Árvore de Decisão, Random Forest, SVM, Regressão Logística e Redes Neurais (Deep Learning)). Depois de aplicados os algoritmos, foi elencado o algoritmo que teve os me-

[3]https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

lhores desempenhos na base de dados, baseados nas métricas de acurácia e matriz de confusão por ele gerado.

5. *Realização de testes com a base de dados para verificação dos atributos que mais influenciam na evasão dos alunos:* Definido o melhor algoritmo, foram feitos testes com o mesmo dentro da base de dados, levando em consideração alguns subconjuntos de atributos, do conjunto total contido na base. Esta técnica foi utilizada para que, de fato, fosse possível concluir com o máximo de exatidão, quais foram os atributos que exerceram mais influência na evasão de alunos no Ensino Superior.

6. *Indicação dos atributos que mais influenciam na evasão de alunos:* Com foco na base de dados investigada, foram descritos os atributos que mais influenciaram na evasão dos alunos.

7. *Criação de um modelo eficiente para previsão da evasão:* Depois de identificar o algoritmo que previu a evasão com mais eficiência e encontrar os atributos que mais influenciaram nesse processo, foi gerado um modelo capaz de prever a evasão de alunos do ensino superior. Objetivou-se que esse modelo pudesse ser utilizado para auxiliar gestores de cursos e Instituições de Ensino a prevenir a evasão em cursos de graduação.

Para a construção desse modelo preditivo, foi necessário seguir um processo estruturado de análise de dados e experimentação com diferentes técnicas de aprendizado de máquina. Inicialmente, a base de dados foi tratada e preparada, garantindo que os atributos mais relevantes fossem considerados. Em seguida, diversos algoritmos foram testados para avaliar seu desempenho na previsão da evasão. Esse processo possibilitou a escolha do modelo mais adequado e sua posterior aplicação na base reformulada, garantindo maior precisão na identificação de padrões de evasão estudantil.

Como é ilustrado na Figura 3, será utilizada uma base de dados já existente, na qual será realizado o pré-processamento dos dados. Com o pré-processamento feito, serão aplicados os algoritmos de aprendizagem de máquina escolhidos para este trabalho. Posteriormente, será elencado o melhor algoritmo dentro dos escolhidos. Após a escolha do melhor, testes serão realizados na base de dados, utilizando conjuntos de atributos específicos, para que se chegue a um conjunto ideal de atributos para predição da evasão dos alunos. Na sequência, um modelo será gerado, utilizando a base de dados reformulada e o algoritmo identificado como mais eficaz. Esse modelo poderá ser um apoio na demanda contra a evasão em cursos de ensino superior. Por fim, os resultados serão documentados e expostos no trabalho.

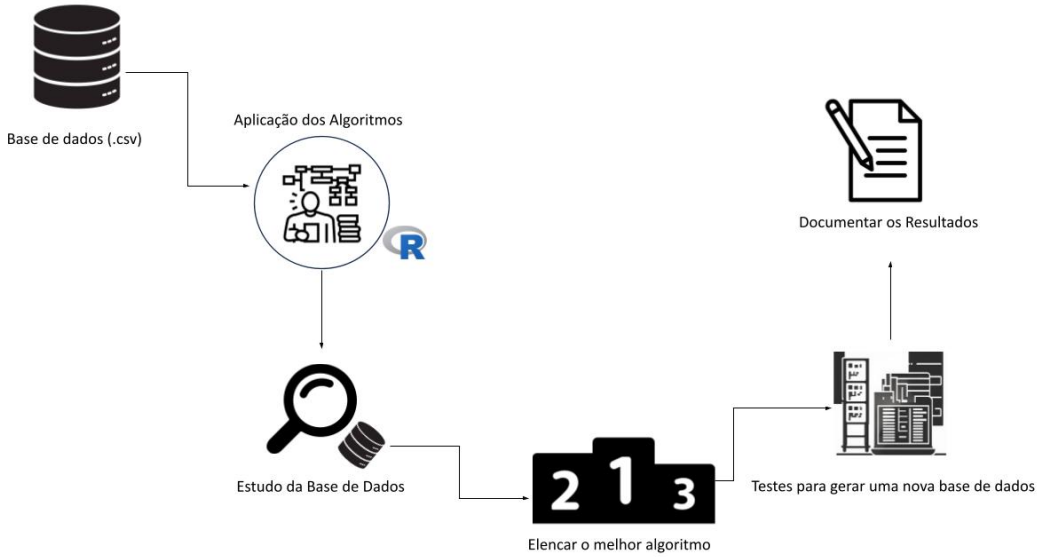

Figura 3. Esboço do Estudo. Fonte: Autor

Cabe destacar, que toda a manipulação da base de dados, bem como a aplicação dos algoritmos de AM foi realizada, por meio da linguagem de programação R. A linguagem R é uma linguagem estatística e gráfica, multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada, voltada à manipulação, análise e visualização de dados. Ela é considerada bastante eficaz para esses fins. A linguagem R foi criada no departamento de Estatística da Universidade de Auckland, Nova Zelândia, por Ross Ihaka e Robert Gentleman, na década de 1990.

A linguagem R é multiplataforma, pode ser executada em diferentes sistemas operacionais, como Windows, Linux e Macintosh, é dinamicamente tipada, orientada a objetos e possui código aberto. Pelo fato de a linguagem R ser amplamente utilizada na manipulação, análise e visualização de dados, muitas vezes ela nem chega a ser considerada como linguagem de programação, mas sim, como um produto estatístico especializado. Convém mencionar que além da linguagem R, existe também o ambiente para computação estatística e gráficos R Studio, que é um conjunto de instalações de software. Para o desenvolvimento deste trabalho, foi realizado um processo de análises de algoritmos de Aprendizagem de Máquina, e de análise da base de dados e dos subconjuntos que esta pode formar, devido à sua grande amplitude (37 atributos). Durante todo o desenvolvimento, foi utilizada a linguagem R, por meio da Interface de Desenvolvimento R Studio.

5. Resultados

Nesta seção são apresentados os resultados deste estudo, primeiramente são apresentadas algumas estatísticas descritivas sobre a base de dados, posteriormente são relatadas as aplicações dos algoritmos na base de dados completa, na sequência são descritas as aplicações do algoritmo com maior acurácia nos subconjuntos da base completa.

5.1. Análise Descritiva dos Dados

A base de dados utilizada para análise neste trabalho trata-se de uma base de dados educacional proveniente do Instituto Politécnico de Portalegre, em Portugal. Base de dados também utilizada no trabalho de Martins et al. (2021). Esta base de dados contém informações sobre 4.424 estudantes universitários, contendo 37 características diferentes de todos os alunos, além da informação do status dos alunos, se eles se formaram, se estão cursando ou se evadiram do curso em que estavam matriculados.

Esta base de dados inclui informações importantes como: dados acadêmicos, dados familiares, dados socioeconômicos e dados pessoais do aluno. No decorrer do trabalho, iremos tratar essas características da base de dados como "atributos". Dentre os 37 atributos da base de dados, vale destacar alguns atributos que são relevantes para a nossa análise como um todo, como a idade média e a média das notas dos alunos. Quanto ao status dos alunos, do total de 4.424 alunos em questão, 2.209 alunos se graduaram (50%), 1.421 evadiram do curso (32%) e 794 ainda estão matriculados(18%). Conforme ilustrado na Figura 4.

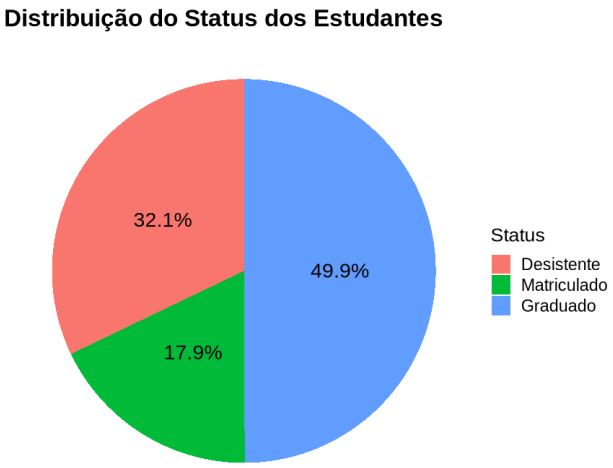

Figura 4. Status dos estudantes. Fonte: Autor

A idade média dos alunos é de 23 anos. Dentre os alunos, 1.099 são estudantes com bolsa (25%) e 3.325 são estudantes sem bolsa (75%). Se tratando da média das notas dos alunos, as médias dos alunos desistentes eram mais baixas, aproximadamente 3.63 no 1º semestre e 2.95 no 2º semestre. Os alunos ainda matriculados possuem uma média intermediária de 5.56 para ambos os semestres. Os alunos que já se graduaram possuem médias mais altas, 6.32 para o 1º semestre e 6.35 para o 2º semestre. Esses dados reforçam a relação entre o desempenho acadêmico do aluno e seu status. Os dados acerca das médias dos alunos estão dispostos na Figura 5.

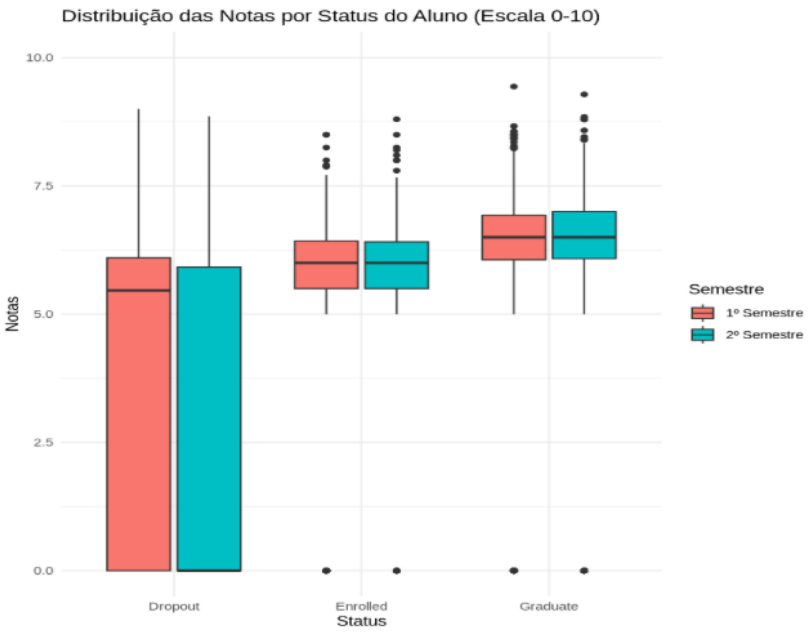

Figura 5. Médias dos Alunos. Fonte: Autor

Ter o conhecimento profundo de todos os atributos desta base é fundamental para a realização do estudo proposto, pois nos permite identificar relações potenciais entre os atributos presentes na base de dados, como, por exemplo, a correlação entre as notas dos alunos e seu status. Com isso, é possível desenvolver hipóteses relevantes baseadas na utilização dos algoritmos citados neste trabalho, podendo garantir a qualidade e a confiabilidade das conclusões obtidas.

5.2. Resultados Experimentais

Primeiramente, a base de dados passou por um processo de pré-processamento dos dados, onde há a verificação se os dados da base necessitam de algum ajuste (se existem dados faltantes, se é necessário fazer escalonamento, se existem valores muito discrepantes, por exemplo), para melhorar ainda mais a precisão dos algoritmos que nela serão aplicados. Na parte do pré-processamento dos dados foi calculada a média das informações para os atributos que continham valores faltantes. Ademais, destaca-se que o método utilizado para avaliar os modelos gerados foi o modelo de treinamento e teste, em que uma parte dos dados é utilizada para treinar e outra para testar o modelo, sendo estes divididos de forma aleatória utilizando uma biblioteca da linguagem R, em todos os experimentos realizados com a base de dados completa e seus subconjunto foram utilizados 80% para treinamento e 20% para teste.

Nesse sentido, foi realizada uma aplicação de todos os algoritmos descritos no trabalho: Naive Bayes, Árvore de Decisão, Random Forest, SVM(Support Vector Machines), Regressão Logística e Redes Neurais na base de dados completa. A partir dessa aplicação foi possível identificar o modelo que melhor se ajustou à base de dados, este posteriormente foi utilizado para dar continuidade ao processo de estudo sobre os atributos dos alunos. Para o estudo da relação dos atributos com a evasão dos alunos a base de dados foi dividida em 4 subconjuntos:

Base A - que engloba as características familiares, é constituída por 5 atributos (colunas);

Base B - que engloba as características pessoais do aluno, é constituída por 6 atributos (colunas);

Base C - que engloba as características socioeconômicas, é constituída por 7 atributos (colunas);

Base D - que engloba as características acadêmicas, é constituída por 19 atributos (coluna);

Depois dessa divisão foi possível aplicar o algoritmo de AM que gerou o modelo mais acurado para cada subconjunto e verificar quais atributos podem ser considerados os melhores preditores da evasão dos alunos em curso de Graduação. Um detalhamento deste processo é destacado na sequência.

5.2.1. Base Completa

Primeiramente foi feita a aplicação de todos os algoritmos citados na base completa, com os 37 atributos. Esta análise foi importante, pois a partir do algoritmo que gerou o modelo mais aderente à base de dados é que se daria sequência à análise, evitando trabalho desnecessário, com algoritmos que não geram modelos eficientes para os dados em questão. Os resultados das métricas de avaliação definidas neste trabalho para os algoritmos, aplicados à base de dados completa estão descritas na Tabela 4.

Tabela 4. Desempenho de Algoritmos de Aprendizado de Máquina

Algoritmo	Configurações	Acurácia	Intervalo de Confiança	TNI	Valor de P	Kappa
Naïve Bayes	Default	0.6942	(0.6627, 0.7245)	0.4994	2.2×10^{-16}	0.5095
Árvore de Decisão	Default	0.7407	(0.7104, 0.7693)	0.4994	2×10^{-16}	0.5717
Random Forest	ntree = 100	0.778	(0.7491, 0.805)	0.4994	2.2×10^{-16}	0.627
SVM	Kernel = Linear	0.773	(0.7495, 0.7953)	0.4992	2.2×10^{-16}	0.6154
Regressão Logística	Multinomial	0.767	(0.7433, 0.7895)	0.4992	2.2×10^{-16}	0.6062
Redes Neurais	5 camadas, 1000 épocas	0.7594	(0.7355, 0.7822)	0.4992	2.2×10^{-16}	0.5861

Como pode ser percebido o modelo gerado pelo algoritmo Random Forest alcançou uma acurácia de 77,8%, este gerou o modelo mais aderente à base de dados, para mais detalhes sobre a implementação é possível consultar os códigos em um repositório público[4]. Dessa forma, a partir deste ponto, as análises nas bases resultantes da divisão da base completa são feitas com este algoritmo apenas. É importante ainda destacar que, como este é um problema de classificação de três categorias, na qual há um forte desequilíbrio em relação a uma das classes, não seria possível chegar a um desempenho muito alto, como em outras bases de dados que são problemas de classificação de duas catego-

[4]https://github.com/felipom/Desenvolvimento-TCC.git

rias, como se vê na maioria dos estudos, que abordam o problema com “evadiu” ou “não evadiu”.

Salienta-se também que, dentre todos os 37 atributos dos alunos, aquele que se mostrou mais relevante foi o que engloba as aprovações nos componentes curriculares do segundo semestre (Curricular.units.2nd.sem.approved.<5 - Figura 6), dependendo da aprovação ou não do aluno nos mesmos. Se este valor ficar abaixo de 5, a probabilidade do aluno evadir aumenta consideravelmente e caso fique acima, é provável que esse aluno venha a terminar o curso. Não se trata de uma regra, mas são especulações plausíveis baseadas nas informações que foram levantadas.

Pode-se validar essa informação na Figura 6, esta é resultante da aplicação do algoritmo de Árvore de Decisão, que é a base da inteligência do algoritmo do Random Forest, o mesmo utiliza a entropia e o ganho de informação para chegar nesta árvore, o que indica que os atributos que a compõem têm um grau de importância elevado, e o atributo que fica no nó raiz é o que tem a maior relevância no processo de predição . Ademais, como foi observado o próprio algoritmo de Árvore de Decisão obteve um bom desempenho sobre os dados, chegando a uma acurácia de aproximadamente 74%.

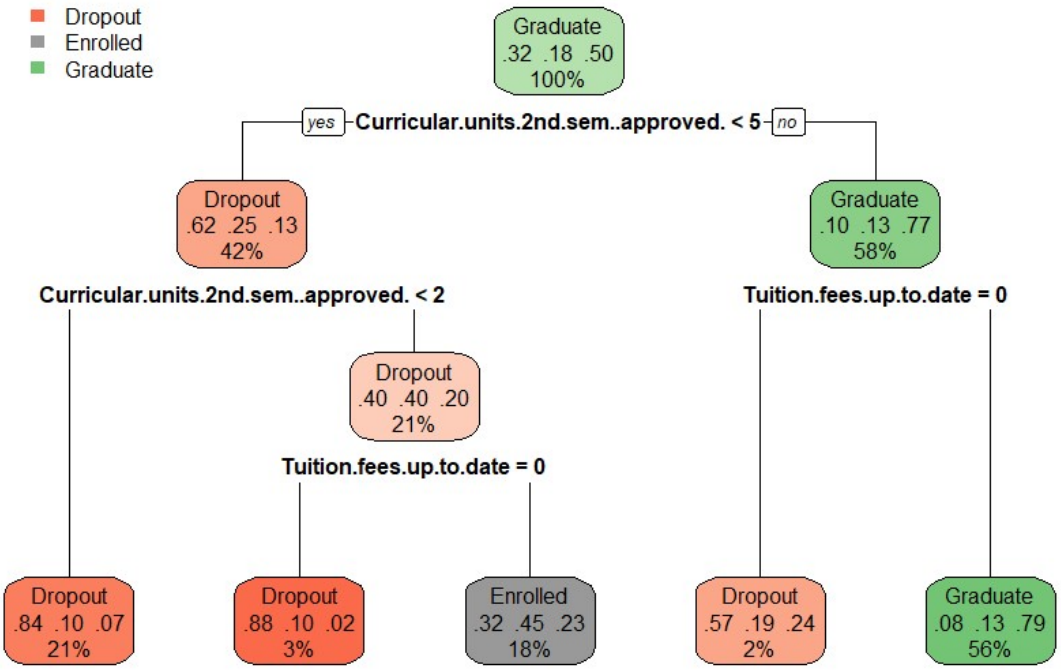

Figura 6. Árvore de decisão para a base completa. Fonte: Autor

5.2.2. Subconjuntos da Base: Aplicações e Resultados

A base de dados completa foi dividida em quatro subconjuntos. Após essa divisão, o algoritmo Random Forest foi aplicado a cada um deles, pois apresentou o melhor desempenho na predição da evasão dos alunos. Esse processo permitiu identificar qual dos subconjuntos gerou o modelo mais acurado. Além disso, a análise das árvores de decisão

resultantes possibilitou a identificação do atributo mais relevante em cada subconjunto. Os resultados da aplicação do algoritmo Random Forest em cada um dos subconjuntos são detalhados na Tabela 5.

Tabela 5. Desempenho do Random Forest nos subconjuntos da base de dados

Base	Configuração	Acurácia	Intervalo de Confiança	TNI	Valor de P	Kappa
Base A - Atributos Familiares	ntree = 100	0.5155	(0.5074, 0.5328)	0.4992	5.2×10^{-16}	0.1227
Base B - Atributos Pessoais	ntree = 100	0.5362	(0.5074, 0.5618)	0.4992	5.3×10^{-16}	0.1617
Base C - Atributos Socioeconômicos	ntree = 100	0.6018	(0.5802, 0.6335)	0.4992	4.4×10^{-16}	0.2584
Base D - Atributos Acadêmicos	ntree = 100	0.7352	(0.7114, 0.7596)	0.4992	2.8×10^{-16}	0.5497

O subconjunto que trata das questões familiares do aluno é formado por 5 atributos: qualificação acadêmica da mãe, qualificação acadêmica do pai, ocupação da mãe, ocupação do pai e se o estudante é deslocado de sua residência. O Random Forest teve uma acurácia de 51%, um intervalo de confiança entre 0.5074 e 0.5328, uma taxa de não informação de 0.4992, valor de p de 0.005213 e o valor do kappa de 0.1227 e o atributo com maior relevância deste conjunto foi a ocupação da mãe, conforme exemplificado na figura 7.

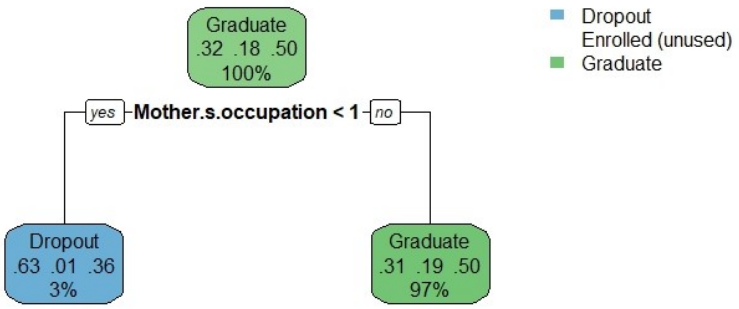

Figura 7. Árvore de decisão para os atributos familiares. Fonte: Autor

O subconjunto dos atributos pessoais é formado pelo estado civil do estudante, nacionalidade, se o estudante tem alguma necessidade educacional especial, gênero, idade em que se matriculou no curso e se ele é estrangeiro ou não. Para este subconjunto, o atributo mais impactante foi a idade em que o aluno se matriculou, como pode ser observado na Figura 8, e o modelo gerado pelo algoritmo de AM, Random Forest, teve uma acurácia de 53%, um intervalo de confiança entre 0.5074 e 0.5618, uma taxa de não informação de 0.4992, valor de p de 0.005313 e o valor do kappa de 0.1617.

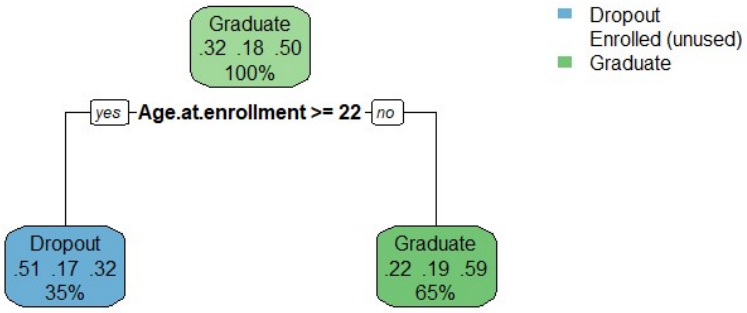

Figura 8. Árvore de decisão para os atributos pessoais. Fonte: Autor

Para o subconjunto das informações socioeconômicas dos estudantes, temos os seguintes atributos: informação de se o estudante é devedor ou não, se suas mensalidades estão em dia, se ele é ou não bolsista, e também temos informações sobre a taxa de desemprego, taxa de inflação e PIB de sua localidade. Para os aspectos socioeconômicos, o atributo que mais se destacou foi acerca das mensalidades estarem em dia do estudante, como destacado na Figura 9. O Random Forest teve uma acurácia de 60% quando aplicado a este subconjunto, um modelo com uma acurácia relevante para uma base de dados com poucos atributos. Além da acurácia, o algoritmo apresentou para o intervalo de confiança entre 0.5802 e 0.6335, taxa de não informação de 0.4992 e para o valor de kappa 0.2584.

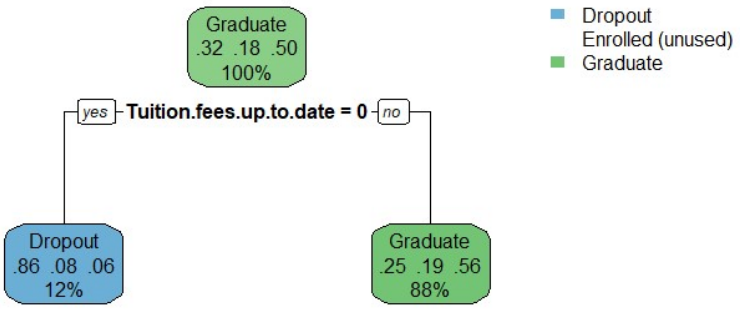

Figura 9. Árvore de decisão para os atributos socioeconômicos. Fonte: Autor

Por fim, o subconjunto dos atributos acadêmicos é composto pelo modo de inscrição, ordem de preferência da inscrição, curso, turno, qualificação anterior do estudante, nota da qualificação anterior, nota de admissão, unidades curriculares creditadas no 1° semestre, unidades curriculares matriculadas no 1° semestre, avaliações no 1° semestre, unidades curriculares aprovadas no 1° semestre, notas do 1° semestre, unidades curriculares sem avaliação no 1° semestre, unidades curriculares creditadas no 2° semestre,

unidades curriculares matriculadas no 2° semestre, avaliações no 2° semestre, unidades curriculares aprovadas no 2° semestre, notas do 2° semestre e unidades curriculares sem avaliação no 2° semestre. Totalizando 19 atributos na base de dados.

Neste subconjunto, o atributo que se mostrou mais preditor foi o das unidades curriculares aprovadas do 2° semestre, como ilustrado na Figura 10. O algoritmo Random Forest teve uma acurácia de 73%, a maior dentre todos os subconjuntos, além de apresentar um intervalo de confiança entre 0.7114 e 0.7596, taxa de não informação de 0.4992, valor de p 0.2873 e o valor de kappa em 0.5497.

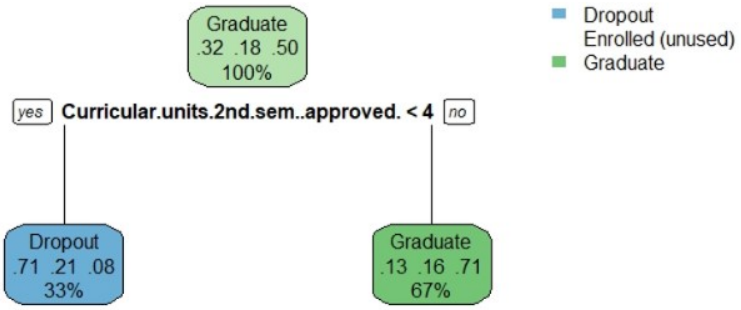

Figura 10. Árvore de decisão para os atributos acadêmicos. Fonte: Autor

6. Discussões

Como esperado com os experimentos realizados, a base que possui os dados mais preditores de evasão dos alunos é a base de dados que diz respeito aos dados acadêmicos, pois estes estão diretamente ligados à evasão ou conclusão dos alunos. Todavia, algumas constatações interessantes foram identificadas no decorrer do processo, estas são discutidas nesta seção.

Primeiramente, é interessante destacar que este estudo utilizou a base de dados sistematizada por Martins et al. (2021), um dos trabalhos citados na seção trabalhos relacionados, e corroborando este estudo, o melhor desempenho identificado por algoritmos não combinados, também foi o modelo gerado pelo algoritmo Random Forest, na base completa (visto que no estudo de Martins et al. (2021) somente foram aplicados os algoritmos na base completa). Foi observado que os autores chegaram a uma acurácia de 72% com o Random Forest, e mesmo utilizando modelos combinados os autores chegaram a 73% de acurácia com o "Extreme Gradient Boosting". Acredita-se, que no estudo aqui desenvolvido foi possível chegar a 78% de acurácia, na base completa, pela forma como a base de dados foi pré-processada, e também pela quantidade de árvores de decisão utilizadas (100 árvores). Desta forma, é possível considerar que o desenvolvimento deste estudo foi conduzido de forma adequada, vide a similaridade com estudos já publicados na área de Mineração de Dados Educacionais.

Neste sentido, destaca-se também que o modelo gerado em Hegde (2018) para prever, de forma precoce, a possível evasão de alunos de seus respectivos cursos, que obteve também uma acurácia similar ao estudo de Martins et al. (2021), e o presente trabalho, alcançando 72% de acurácia. Demonstrando que o estudo aqui desenvolvido obteve um ganho significativo em termos de acurácia no modelo de previsão da evasão, em se comparando com um pequeno recorte no estado da arte.

No que tange ainda aos trabalhos relacionados, cabe destacar que o estudo desenvolvido por Souza (2020), no qual foi realizada a previsão do desempenho de alunos realizando um comparativo entre as técnicas consolidadas na área da Mineração de Dados Educacionais e a técnica de Aprendizagem Profunda, apresentou uma forma interessante de analisar os atributos mais relevantes de uma base de dados, que é por meio da geração da árvore de decisão. Essa metodologia foi adotada neste estudo, pois por meio das métricas utilizadas para a geração da árvore é possível dizer que o atributo do nó raiz e os mais próximos dele são os mais relevantes para prever o atributo meta (se o aluno evadiu), em meio a todos os atributos da base de dados.

Em seguida, destaca-se alguns fatores importantes identificados quando se prossegue no estudo da base de dados, indo além da geração do modelo de predição da evasão, que é o estudo sobre os atributos que mais influenciam neste processo. Para isso, a base de dados foi dividida em 4 subconjuntos de dados que se relacionam, com: características familiares, características pessoais, características socioeconômicas e características acadêmicas; como observado na seção anterior.

A partir disso foram feitas análises para descobrir dentro de cada subconjunto qual o atributo era o mais relevante. Na base com dos dados familiares foi identificado que o trabalho da mãe é um fator de grande relevância para permanência dos alunos. Concluise que, dependendo do trabalho da mãe, as chances de um alunos evadir ou não, podem variar de acordo.

No que se refere, às características pessoais dos alunos, um fator decisivo é a idade, sendo que alunos com 24 anos ou mais têm maior probabilidade de se formar, em detrimento a alunos com menos idade. É de conhecimento que a idade para um jovem entrar em um curso de graduação é baixa, e muitas vezes esses alunos não têm conhecimento, ou a vivência necessária para saber qual é realmente a carreira que querem seguir, isso é um fator que impacta na evasão dos alunos.

Quanto aos atributos que correspondem a aspectos socioeconômicos, destaca-se o pagamento das mensalidades, como os cursos analisados não são gratuitos, este é um aspecto de grande relevância, pois se o aluno não tem condições financeiras para realizar os pagamentos das mensalidades, há uma forte tendência à evasão. Destaca-se ainda que, a base com os atributos socioeconômicos foi a que teve o segundo melhor modelo para predição da evasão dos alunos, chegando a 60% de acurácia, dessa forma, percebe-se que as condições financeiras dos alunos impactam fortemente na decisão de continuar ou não em curso de graduação.

Por fim, como já era esperado os atributos que melhor predizem se o aluno tem propensão a evadir ou não são os atributos acadêmicos, com um modelo que alcançou 73% de acurácia. Como pode ser observado nas unidades curriculares aprovadas do 2° semestre, são um forte indicativo se o aluno irá evadir ou não, se o aluno tiver mais de

4 aprovações ele está mais propenso a concluir que evadir o curso. Então a partir do 2° semestre já é possível identificar alunos propensos à evasão, e realizar um trabalho para permanência destes alunos.

Em suma, destaca-se que é possível a partir das análises desenvolvidas neste estudo, elaborar uma base de dados geral que possa ser adequada para mais instituições de ensino superior, englobando atributos (características) dos alunos que possam predizer a evasão mais no ínicio do curso, auxiliando gestores educacionais, coordenadores pedagógicos e de cursos a realizar intervenções antes que o aluno evada, ou que a situação do aluno fique ruim a ponto que este se sinta desmotivado a concluir.

7. Considerações Finais

O objetivo geral deste estudo foi o de identificar quais são os principais atributos que implicam na evasão de alunos do Ensino Superior, por meio de técnicas da mineração de dados educacionais com ênfase a aprendizagem de máquina. Por meio das análises realizadas, os resultados demonstraram que o objetivo proposto foi alcançado. Foi possível não apenas segmentar a base de dados em quatro subconjuntos distintos, mas também determinar, com um nível de acurácia satisfatório, qual desses subconjuntos exerce a maior influência sobre a taxa de evasão dos estudantes. Dessa forma, pode-se responder a questão de pesquisa levantada: Quais são as variáveis que mais impactam na evasão de um aluno no ensino superior? Primeiramente, são as variáveis acadêmicas em que foi percebido que para os alunos em questão, um semestre crucial para saber se iriam completar ou não o curso era o 2° semestre. Se os alunos tivessem mais de 4 aprovações neste semestre, a probabilidade de evasão era menor. Além disso, outras variáveis importantes foram identificadas como: a formação da mãe, a quantidade de mensalidades atrasadas e a idade.

Cabe destacar, que apesar do enfoque do estudo estar em identificar os atributos mais relevantes para predição da evasão, foi necessária a geração de modelos que fizessem essa previsão. Depois da aplicação de vários algoritmos de Aprendizagem de Máquina sobre a base de dados, os testes realizados indicaram que o algoritmo Random Forest obteve o melhor desempenho. O modelo gerado na base completa obteve 78% de acurácia; o modelo gerado a partir dos atributos familiares alcançou 51%; o modelo gerado a partir da aplicação do algoritmo na base dos atributos pessoais teve 53% de acurácia; dos atributos socioeconômicos teve 60%; e por fim dos atributos acadêmicos alcançou 73% de acurácia.

Dessa forma, fica evidente que o estudo alcançou as metas estabelecidas, as quais eram: Analisar uma base de dados de alunos do ensino superior; Aplicar e avaliar algoritmos de aprendizagem de máquina (Naive Baïes, Árvore de Decisão, Random Forest, Support Vector Machines (SVM), Regressão Logística e Redes Neurais) na perspectiva de selecionar o mais eficaz; Realizar testes sobre a base de dados, subdividindo a mesma, para verificar qual o conjunto de atributos que mais impactam na evasão do aluno; Selecionar e descrever os atributos identificados como os mais significativos na evasão dos alunos; Gerar um modelo eficaz para a predição da evasão de alunos no ensino superior.

Em suma, este estudo apresenta uma contribuição significativa, ao fornecer uma visão estruturada e baseada em dados sobre os fatores determinantes para a evasão no Ensino Superior. Com a determinação de forma precoce de tendências de evasão é possí-

vel que sejam realizadas intervenções na perspectiva de apoiar os alunos a concluírem os cursos.

Como trabalhos futuros, pretende-se montar uma base de dados com os atributos mais relevantes identificados neste estudo e realizar a aplicação dos algoritmos para verificar se é possível chegar a indicadores mais precisos da evasão dos alunos, com a menor quantidade de de atributos possível e com o menor tempo de entrada dos alunos nos cursos. Para assim, identificar tendências de evasão logo no início para que as intervenções a serem realizadas sejam mais efetivas, criando uma cultura de permanência e êxito.

Referências

AGGARWAL, C. C. *Data Mining: The Textbook*. 2015. Disponível em: <https://doi.org/10.1007/978-3-319-14142-8>. Último acesso em: 26/11/2023.

ALVES, M. de O. P.; GAYDEZKA, B.; CAMPOS, A. de. Projeto para registro e controle da evasão na uftm. v. 11, n. 1, p. 125–135, 2018.

BAKHSHINATEGH, B. et al. *Education and Information Technologies*. 2018. Disponível em: <https://doi.org/10.1007/s10639-017-9616-z>. Último acesso em: 26/11/2023.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 2006.

EDM. *Educational Data Mining*. 2020. Disponível em: <http://educationaldatamining.org>. Último acesso em: 26/11/2023.

FILHO, R. L. L. e S. et al. A evasão no ensino superior brasileiro. v. 37, n. 132, p. 641–659, 2007.

FRITSCH, R.; ROCHA, C.; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. v. 52, n. 38, p. 81–108, 2015.

GIL, C. A. *Como elaborar projetos de pesquisa*. 4ª. ed. São Paulo, SP: Atlas, 2002.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2ª. ed. California, USA: Springer, 2009.

HEGDE, V. *Higher Education Student Dropout Prediction and Analysis through Educational Data Mining*. 2018.

JAPKOWICZ, N.; SHAH, M. *Evaluating Learning Algorithms: A Classification Perspective*. 1ª. ed. Cambridge, UK: Cambridge, 2014.

LOBO, M. B. C. de M. *Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções Associação Brasileira de Mantenedoras de Ensino Superior*. 2012.

MARTINS, M. V. et al. *Early Prediction of student's Performance in Higher Education: A Case Study*. 2021.

MELLO, S. P. T. de et al. *O fenômeno evasão nos cursos superiores de tecnologia: um estudo de caso em uma universidade pública no sul do Brasil.* 2013. Disponível em: <https://repositorio.ufsc.br/handle/123456789/113096>. Último acesso em: 26/11/2023.

MITCHELL, T. M. *Machine Learning*. 1ª. ed. Nova York, USA: MacGraw-Hill, 1997.

NAVARRO, P. J. et al. *A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3D range data*. 2017. Disponível em: <https://doi.org/10.3390/s17010018>. Último acesso em: 27/11/2023.

RISTOFF, D. A tríplice crise da universidade. v. 4, n. 3, p. 9–14, 2014.

SOUZA, V. F. de. *Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda*. 2020.

SOUZA, V. F. de. *Os avanços da mineração de dados educacionais: processo, tendências temáticas e técnicas de mineração*. 1ª. ed. Curitiba, PR: Bagai, 2021.