Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Londrina

# TÉCNICAS DE CLASSIFICAÇÃO PARA A PREDIÇÃO DA EVASÃO UNIVERSITÁRIA

Londrina

2021

**ISADORA GONÇALVES TOLEDO**

# TÉCNICAS DE CLASSIFICAÇÃO PARA A PREDIÇÃO DA EVASÃO UNIVERSITÁRIA

Trabalho de Conclusão de Curso apresentado no curso de graduação em Engenharia de Produção da Universidade Tecnológica Federal do Paraná – Campus Londrina.

Orientador: Dr. Bruno Samways dos Santos

Londrina

2021

Ministério da Educação
**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**
DEP. ACAD. DE ENG. DE PRODUCAO - LD

## TERMO DE APROVAÇÃO

### TRABALHO DE CONCLUSÃO DE CURSO - TCC

### TÉCNICAS DE CLASSIFICAÇÃO PARA A PREDIÇÃO DA EVASÃO UNIVERSITÁRIA

Por

**ISADORA GONÇALVES TOLEDO**

Monografia apresentada às 14 horas 00 min. do dia 13 de Maio de 2021 como requisito parcial, para conclusão do Curso de Bacharelado em Engenharia de Produção da Universidade Tecnológica Federal do Paraná, Câmpus Londrina. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação e conferidas, bem como achadas conforme, as alterações indicadas pela Banca Examinadora, o trabalho de conclusão de curso foi considerado APROVADO.

Banca examinadora:

| | |
|---|---|
| Prof. Dr. Rogério Tondato | Membro |
| Profa. Dra. Silvana Rodrigues Quintilhano Tondato | Membro |
| Prof. Dr. Bruno Samways dos Santos | Orientador |
| Profa. Dra. Silvana Rodrigues Quintilhano Tondato | Professor(a) responsável TCCII |

Documento assinado eletronicamente por (Document electronically signed by) **ROGERIO TONDATO, PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 13/05/2021, às 17:00, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por (Document electronically signed by) **SILVANA RODRIGUES QUINTILHANO TONDATO, PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 13/05/2021, às 17:04, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por (Document electronically signed by) **BRUNO SAMWAYS DOS SANTOS, PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 13/05/2021, às 17:44, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site (The authenticity of this document can be checked on the website) https://sei.utfpr.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador (informing the verification code) **2016064** e o código CRC (and the CRC code) **89802C2A**.

**RESUMO**

Este trabalho busca realizar a predição da classificação de alunos quanto a evasão por meio da aplicação de técnicas de Aprendizado de Máquina. Para isso, foi realizado uma fundamentação teórica sobre as possíveis causas da evasão estudantil e a obtenção de um conjunto de dados através do próprio sistema da universidade. Primeiramente, no Experimento 1, utilizou-se a técnica de Árvore de Decisão, k-NN e RNA, e em seguida, no Experimento 2, utilizou-se a técnica de k-NN e RNA, com atributos reduzidos, selecionados pela Árvore de Decisão no Experimento 1. Foi possível concluir que a técnica mais adequada para realizar a predição da evasão de alunos foi a técnica RNA, com taxa de aprendizagem de 0,1 e com seleção de atributos, que obteve o melhor desempenho em acurácia, com 92,3%, em precisão, com 89,8% e em sensibilidade, com 82,6%. Também foi possível constatar que os principais fatores que podem influenciar na evasão são as questões relacionadas ao desempenho acadêmico dos alunos.

**Palavras-chave:** Aprendizado de Máquina. Predição. Classificação. Evasão.

**ABSTRACT**

This research pursues to predict the classification of students regarding dropout through the application of Machine Learning techniques. For this, a theoretical grounding was made on the possible causes of student dropout and the obtaining of a set of data through the university system itself. First, in Experiment 1, the Decision Tree, k-NN, and RNA technique were used, and then, in Experiment 2, the k-NN and RNA technique was used, but with reduced attributes instead, selected by the Decision Tree in Experiment 1. It was possible to conclude that the most appropriate technique for predicting student dropout was the RNA technique, with a learning rate of 0.1, with selection of attributes, which obtained the best performance presenting accuracy of 92,3%, precision of 89.8%, and recall of 82,6 %. It was also possible to verify that the main factors that can influence dropout are issues related to students's academic performance.

**Keywords:** Machine Learning. Prediction. Classification. Dropout.

## SUMÁRIO

# 1. INTRODUÇÃO

A Mineração de Dados é uma etapa do processo de Descoberta de Conhecimento (*Knowledge Discovery in Databases* – KDD), que buscar encontrar informações relevantes através da utilização de algoritmos para análise de um conjunto de dados (FAYYAD *et al.* 1996)

Estes métodos vêm sendo utilizados para descoberta em diversas áreas do conhecimento, como na área de ciências biológicas, saúde e também na área da educação.

De acordo com Baker *et al.* (2011), a "mineração de dados educacionais", ou EDM, do inglês "*Educational Data Mining*", é uma nova área de pesquisa, onde o foco são os dados da educação, e vem crescendo cada vez mais dentro da comunidade científica. Por outro lado, o autor afirma que no Brasil há poucos trabalhos científicos voltados para o tema.

Segundo Carniel (2013), as IES, possuem um histórico de dados sobre seus alunos, gerados desde a matrícula do aluno, com informações pessoais e socioeconômicas, até informações que são geradas a partir do início da vida acadêmica. Para Pinheiro (2008), *apud* Carniel (2013), esses sistemas de informações possuem uma alta capacidade de armazenamento, gerando um alto volume de informações, sendo muito difícil sua interpretação e dificultando a tomada de decisão. Ao se aplicar o devido tratamento nesses conjuntos de dados, essas informações podem contribuir nas estratégias de retenção de alunos no ensino superior, a fim de reduzir os índices de evasão estudantil.

A evasão estudantil é quando o aluno decide abdicar de seu curso, independentemente dos motivos que levaram a essa decisão. Essas evasões são consideradas uma perda que precisa ser estudada, sendo necessário não somente para medir quantos alunos estão saindo da universidade antes da graduação, mas principalmente para analisar as causas dessas desistências. (LOBO, 2012).

Segundo os estudos de Silva Filho *et al.* (2007), a evasão estudantil é um problema que atinge todos os níveis de ensino, mas, principalmente no Ensino Superior, a evasão é um problema internacional que pode afetar o sistema educacional de todo o país.

No entanto, ainda de acordo com os estudos de Silva Filho *et al.* (2007, p. 2): "são raríssimas as IES brasileiras que possuem um programa institucional

profissionalizado de combate à evasão, com planejamento de ações, acompanhamento de resultados e coleta de experiências bem-sucedidas".

Para a realização dos estudos de previsão tanto da situação acadêmica dos quanto da evasão dos alunos nas Instituições de Ensino Superior (IES) é possível utilizar técnicas de mineração de dados. No entanto, a utilização dessas técnicas no campo estudantil é uma área de pesquisa em expansão, sendo necessário pesquisas complementares para definição de quais atributos são preferíveis para este tipo de pesquisa e quais as técnicas de mineração de dados serão utilizadas (DEKKER *et al.*, 2009).

Neste contexto, o presente trabalho apresenta o seguinte problema: Como extrair informações relevantes sobre evasão universitária a partir de técnicas de mineração de dados?

## 1.1. Objetivo Geral

Aplicar técnicas de mineração de dados para classificação e análise de atributos no contexto da evasão em uma universidade pública do Norte do Paraná.

## 1.2. Objetivos Específicos

1. Compilar os fatores de relevância para a questão da evasão universitária;
2. Coletar dados referentes às condições sociais e acadêmicas dos alunos envolvidos no estudo;
3. Pré-processar o conjunto de dados para identificação e ação sobre valores ausentes e anomalias;
4. Aplicar as técnicas de mineração de dados para a classificação e análise de atributos;

## 1.3. Justificativa

A elevada quantidade de alunos que evadem dos cursos de graduação nas Instituições de Ensino Superior, representa vários fatores de desperdícios econômicos, como descreve Silva Filho *et al.* (2007). No setor público perde-se investimento de recursos, no setor privado é uma perda de receita, além de significar também um desperdício para a sociedade, que não tem o retorno dos profissionais que poderiam estar entrando para o mercado de trabalho.

Os estudos de Silva Filho *et al.* (2007) também apontam que a evasão nas IES no Brasil a um nível macroscópico tem alguma correlação com fatores socioeconômicos, por isso há uma necessidade de realizar estudos com o objetivo de combater as taxas de evasão, evitando os desperdícios, tanto sociais quanto financeiros.

Assim, conforme Hoed (2016) afirma em seus estudos, é de interesse da Universidade identificar as causas dessas desistências para que seja possível melhorar a gestão escolar a fim de capacitar funcionários e professores para que estejam preparados para lidar com o problema da evasão e para que estes alunos tenham melhores chances de continuar na graduação.

Justifica-se essa pesquisa para identificar os principais fatores que podem influenciar a evasão e contribuir com a Universidade para que esta consiga ter um acompanhamento desses alunos e com isso reduzir os índices de evasão.

## 1.4. Estrutura do Trabalho

O trabalho apresentado foi dividido em cinco sessões (incluindo esta seção introdutória. Na segunda seção foi realizado o referencial teórico, onde foi feito um levantamento dos estudos a respeito da evasão escolar e dos métodos de aprendizado de máquina utilizados no trabalho, bem como os trabalhos correlatos ao tema.

Já na terceira sessão foi apresentada a metodologia aplicada no trabalho, descrevendo sobre como foi todo o processo da pesquisa, desde a coleta dos dados até a aplicação das técnicas discutidas.

A quarta sessão apresentou os resultados dos experimentos e comparou o desempenho das técnicas aplicados para diferentes conjuntos e também variações dos parâmetros das técnicas aplicadas

E, por último, na quinta sessão foi feito uma análise dos resultados obtidos e da relevância dos mesmos, além de sugestões de melhorias para trabalhos futuros.

# 2. REFERENCIAL TEÓRICO

O presente capítulo busca conceituar o problema da evasão universitária e a metodologia KDD para a descoberta de conhecimento em bases de dados, bem como suas técnicas e métodos de validação e avaliação.

## 2.1. KDD (*Knowledge Discovery in Databases*)

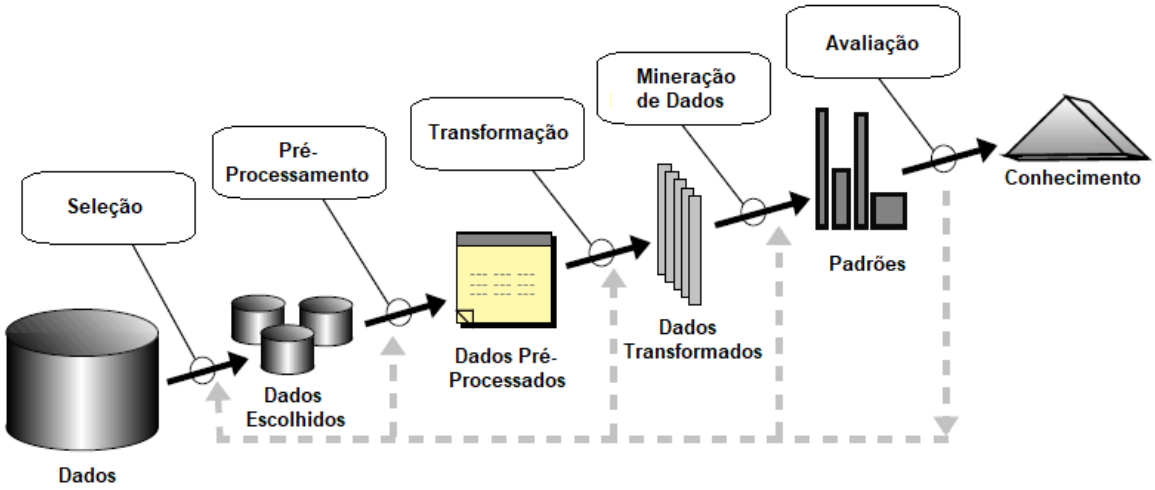
De acordo com os estudos de Fayyad *et al.* (1996), os computadores possibilitaram que os seres humanos conseguissem captar uma grande quantidade de dados e em um ritmo muito acelerado, e devido a essa grande quantidade de dados, maior do que o possível de processar e analisar manualmente, como era realizado no método tradicional.

Como esses métodos tradicionais de interpretação simples tem-se tornado obsoletos, houve a necessidade da criação de abordagens, tecnologias e ferramentas para auxiliar na extração de informações úteis, conhecido como *Knowledge Discovery in Databases* (KDD), que seria a descoberta de conhecimento em bancos de dados.

O KDD está focado em resolver esse grande problema que é a sobrecarga de dados, através do desenvolvimento de métodos e técnicas para auxiliar no entendimento dos dados coletados, como o mapeamento de dados que são muito volumosos para um entendimento fácil.

Ainda de acordo com Fayyad *et al.* (1996), o processo KDD compreende muitas etapas, como mostra a Figura 1, onde primeiramente deve-se obter uma compreensão da aplicação e também entender qual o objetivo do processo do KDD de acordo com o ponto de vista do cliente.

**Figura 1 – Etapas do *Knowledge Discovery in Databases* (KDD)**




Fonte: Adaptado de Fayyad *et al.* (1996)

Em seguida, na segunda etapa do processo, deve-se realizar uma seleção dos dados para criação de um conjunto de dados de destino.

Na terceira etapa ocorre o pré-processamento de dados, onde ocorre a limpeza dos dados. Essa etapa é necessária para preparar a base de dados para a mineração dos mesmos, para isso, pode-se excluir dados faltantes ou incompletos, variáveis irrelevantes para o estudo, integrar ou transformar algum parâmetro, ou seja, é fundamental para adaptar o conjunto de dados para dar continuidade no processo.

A quarta etapa é a redução de dados e projeção, onde devem ser encontrados recursos úteis para representar os dados.

Na quinta etapa, deve-se combinar os objetivos do processo do KDD encontrados na etapa 1 com algum método específico de mineração de dados. A seguir, na sexta etapa é onde ocorre a seleção de modelos e hipóteses, ou seja, quais algoritmos de mineração de dados serão utilizados, bem como quais técnicas serão utilizadas para pesquisa de padrões de dados.

A sétima etapa é quando ocorre de fato a mineração de dados e na oitava, ocorre a interpretação dos padrões encontrados na etapa de mineração de dados, com a possibilidade de retornar a qualquer etapa anterior para mais iterações.

E por último, a nona etapa atua com base nas informações e conhecimento que foram descobertos, seja usando essas informações diretamente, ou aplicando as informações em algum outro sistema ou ainda, apenas documentando e relatando esse conhecimento obtido para as partes interessadas.

## 2.2. Mineração de Dados

Segundo Witten e Frank (2005), a quantidade de dados gerados no mundo não pára de aumentar e a cada dia fica mais fácil armazená-los em formato digital. Desta forma, a mineração de dados é fundamental, pois é um processo que visa encontrar informações relevantes por meio da análise de padrões nesses dados.

Para Fayyad *et al.* (1996), enquanto o termo KDD representa todo o processo de descoberta de conhecimento útil a partir de dados, o termo Mineração de Dados (*Data Mining* – DM) refere-se à uma etapa específica deste processo na qual ocorre a aplicação de algoritmos específicos que funcionam com o objetivo de extrair padrões de dados, envolvendo a montagem de modelos ou determinação de dados observados. Pode haver dois tipos de objetivos para esta etapa do KDD, a verificação, onde o sistema consegue verificar a hipótese do usuário, e a descoberta, onde o sistema encontra novos padrões.

De acordo com McCue (2006), os algoritmos de modelagem são divididos em técnicas de aprendizado de máquina (*Machine Learning* – ML) supervisionadas e não supervisionadas. O objetivo do aprendizado supervisionado é estabelecer regras de decisão que possam ser usadas para determinar um resultado conhecido, incluindo nesta abordagem os modelos de classificação e regressão. O que difere o aprendizado de máquina supervisionado do não-supervisionado é que este último não necessita de uma pré-categorização para os registros (saída ou *output*), e inclui modelos de agrupamento e associação, os quais são usados para agrupamento de dados com atributos relacionados ou similares.

## 2.3. Técnicas de Mineração de Dados

Os objetivos das técnicas de aprendizado de máquina supervisionado e não-supervisionado podem ser alcançados através de métodos de mineração de dados específicos:

a) *Classificação*: busca mapear um determinado registro em uma classe predefinida. Para isso, esse modelo analisa todo o conjunto de dados fornecidos, onde cada registro contém a sua classe para que o modelo consiga “aprender” como classificar um registro novo. (WEISS e KULIKOWSKI, 1991; HAND, 1981, *apud* FAYYAD *et al.*, 1996)

b) *Regressão*: a função mapeia um item em uma variável de valor numérico em vez de um valor categórico, ou seja, é possível estimar qual o valor de uma variável com base nos valores de outras variáveis. (FAYYAD *et al.*, 1996; LAROSE, 2005, *apud* SILVA, 2009)

c) *Clusterização ou agrupamento*: é uma tarefa descritiva com o objetivo de identificar e aproximar os registros com características similares, ou seja, criando uma conjuntos de categorias para descrever os dados (JAIN e DUBES, 1988, TITTERINGTON, SMITH E MAKOV, 1985, *apud* FAYYAD *et al.,* 1996). O que difere da classificação, visto que não é necessário que esses registros sejam pré-categorizados (aprendizado não-supervisionado) e também não tem o objetivo de classificar ou estimar o valor da variável, apenas identificar os grupos de dados similares.

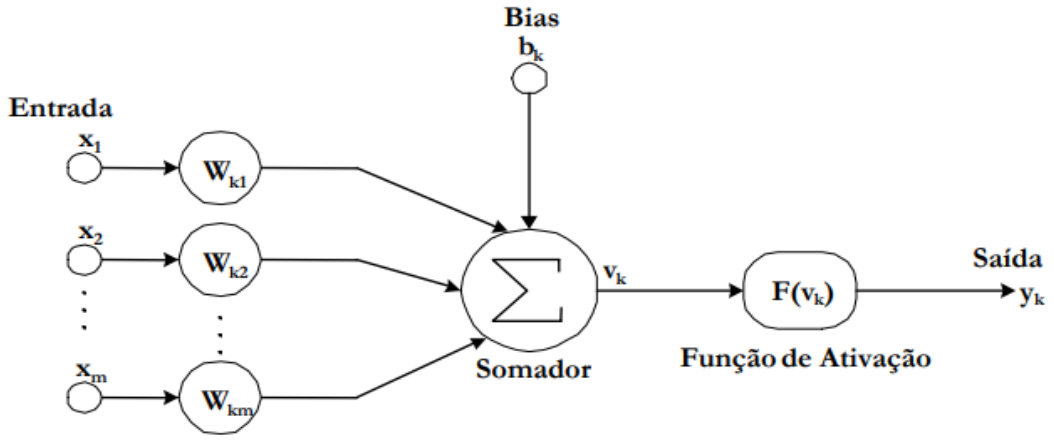
d) *Associação*: é a tarefa que tem o objetivo de encontrar quais atributos estão relacionados, apresentando bons resultados e por isso é uma das tarefas mais conhecidas. (FAYYAD *et al.*, 1996; LAROSE, 2005, *apud* SILVA, 2009)

O presente trabalho irá utilizar o aprendizado de máquina supervisionado com o intuito de classificar a evasão universitária em uma universidade no norte do Paraná, com a aplicação das técnicas de classificação Redes Neurais Artificiais, Árvore de Decisão e k-*Nearest Neighbor*.

### 2.3.1. Redes Neurais Artificiais (RNA)

De acordo com os estudos de Ferneda (2006) as redes neurais artificiais funcionam com o objetivo de simular o cérebro humano, ao tentar reconhecer padrões entre os dados processados através de sua capacidade de armazenar e utilizar conhecimento. Segundo Castro e Zuben (2014) uma RNA pode ser entendida como uma estrutura de processamento, ou seja, uma rede, composta por unidades interconectadas, os neurônios artificiais, onde cada uma dessas unidades apresenta um comportamento específico de entrada/saída. A Figura 2 apresenta um modelo de uma rede neural:

**Figura 2 – Modelo do Neurônio Artificial**




Fonte: Oliveira (2005)

Esses neurônios são fundamentais para a operação das redes neurais e seus principais elementos básicos são, de acordo com Oliveira (2005):

- Sinapse: é uma conexão com um “peso”. Esse peso irá estabelecer a importância dos sinais de entrada. Ou seja, a sinapse está conectada a um neurônio e irá receber um sinal, esse sinal será multiplicado pelo peso sináptico. As sinapses podem ser inibitórias (valor do peso negativo) ou excitatórias (valor do peso positivo)
- Somador: é o elemento que irá somar todas as entradas, já ponderadas pelos pesos sinápticos, de um neurônio
- Função de Ativação: é o que define o sinal de saída de um neurônio de acordo com o valor da soma ponderada das entradas.

Ainda segundo os estudos de Oliveira (2005), as redes neurais podem ser constituídas por uma ou mais camadas. Uma camada é um “agrupamento de neurônios que recebem informações simultaneamente”, ou seja, cada agrupamento representa uma camada de uma RNA.

Essas camadas ainda podem ser classificadas em pelo menos três tipos: camadas de entrada, camadas ocultas e camadas de saída. A camada de entrada não realiza nenhum procedimento, ela é quem apresenta os padrões de entrada à rede neural. A camada oculta tem ligação com as camadas de entrada e de saída, ela

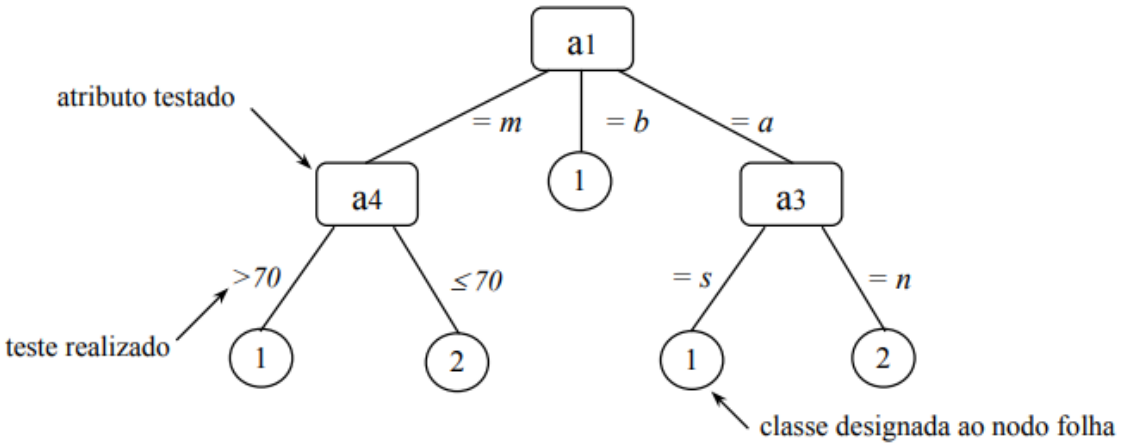
é responsável por extrair características do ambiente. E por fim, a camada de saída é responsável por construir os padrões de resposta da RNA.

### 2.3.2. Árvores de Decisão (AD)

Segundo Shiba *et al.* (2005), as árvores de decisão são utilizadas como algoritmos de classificação e podem ser definidas como representações simplificadas do conhecimento adquirido. De acordo com Garcia (2003), a base desses algoritmos é a estratégia de “dividir para conquistar”, isso quer dizer que o objetivo desses classificadores é fazer com que um conjunto de dados seja dividido em subconjuntos, formados a partir de suas características em comum.

Ainda de acordo com os estudos de Garcia (2003), cada subconjunto obtido nas divisões é conhecido como nodo ou nó, e a classificação final obtida pelo caminho percorrido através desses nós é conhecida como folha. A Figura 3 mostra a representação de uma árvore de decisão com três atributos selecionados e duas classes.

**Figura 3 – Exemplo de um classificador utilizando árvore de decisão**




Fonte: Garcia (2003)

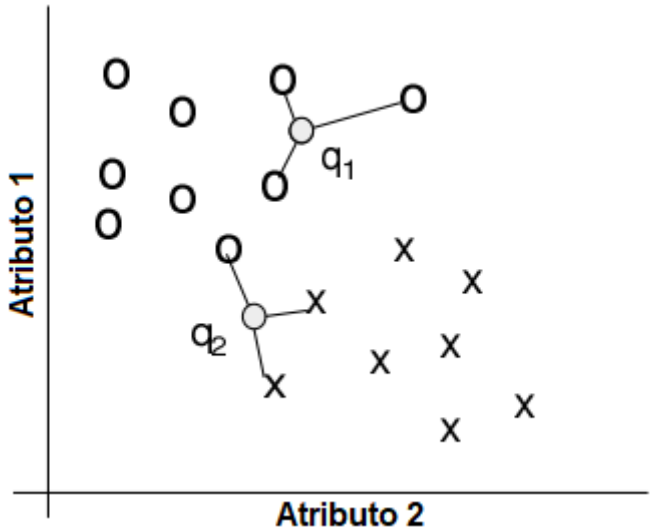
Os nós são os pontos onde são realizados os testes lógicos para a separação de dados. O nó a1 é chamado de nó raiz e é o atributo principal da rede, assim, pode-se dizer que quanto mais alto for a localização de um determinado atributo na árvore de decisão, maior a sua importância. Os nós localizados abaixo do nó a1 são chamados de nós-filhos e estão conectados por ramos. Os valores atribuídos às classificações desses nós-filhos são chamados de folhas.

### 2.3.3. K-*Nearest Neighbors* (k-NN)

Para Cover e Hart (2018) a técnica de $k$-vizinhos mais próximos ($k$-*nearest neighbors* - k-NN) é a decisão não paramétrica mais simples existente, ele consegue atribuir à amostra de entrada não classificada a uma classe da amostra mais próxima no conjunto de treinamento.

Segundo Cunningham e Delany (2007) os exemplos são classificados com base na proximidade de seus vizinhos mais próximos e a maioria das vezes é mais indicado considerar mais de um vizinho, por isso, esse algoritmo é conhecido como k-NN, ou seja, $k$-vizinhos mais próximos, pois o $k$ é o número de vizinhos a ser usado na decisão da classe. Para Cunningham e Delany (2007), "a classificação k-NN tem dois estágios; a primeira é a determinação dos vizinhos mais próximos e a segunda é a determinação da classe usando esses vizinhos". A Figura 4 ilustra este procedimento.

**Figura 4 – Exemplo de k-NN para um valor de $k$ = 3**




Fonte: Adaptado e traduzido de Cunningham e Delany (2007)

Observando a Figura 4, tem-se o exemplo de um modelo de classificação k-NN, onde o número de $k$ considerado é 3. Assim, é possível observar que para $q_1$, os três vizinhos mais próximos são da classe *O*, então $q_1$ é classificado como *O*. Para $q_2$, a classificação é diferente, pois nesse caso $q_2$ tem 2 vizinhos da classe $X$ e um vizinho da classe *O*, portanto, a classificação pode ser resolvida por votação majoritária simples ou votação ponderada a distância. Existem alguns métodos comuns para o

cálculo dessa distância, como a distância Manhattan e a distância Minkowski, mas o mais utilizado é a distância Euclidiana, dada pela Equação (1):

$$d\,(x,y) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \cdots + w_n(x_n - y_n)^2} \quad (1)$$

Onde:

$n =$ número de atributos

$w_i =$ peso relacionado ao atributo $i$

$x_i =$ valor do atributo $i$ para a distância $x$

$y_i =$ valor do atributo $i$ para a distância $y$

## 2.4. Métodos de validação e avaliação

Uma grande quantidade de dados não é garantia de um bom funcionamento do algoritmo, por isso é necessário a utilização de métodos de validação, para garantir que os resultados que o algoritmo apresenta são confiáveis. Para isso, também é necessário analisar as métricas de avaliação, como acurácia, sensibilidade e precisão.

### 2.4.1. Validação de Modelos

Os métodos de validação e avaliação buscam aumentar a confiabilidade da precisão e acurácia de seus resultados, onde, precisão é a porcentagem de acertos de resultados verdadeiro positivo dentre todos os resultados positivos (verdadeiro positivo e falso positivo), e acurácia é a porcentagem de acerto de todos os resultados verdadeiros dentre todos os resultados positivos e negativos.

Os principais métodos de validação são *k-fold* e *hold-out*, que serão definidas a seguir:

a) Método $k$*-fold*

De acordo com Duchesne *et al.* (2005), conforme explica Schreiber (2017), a validação $k$*-fold* é um método de validação cruzada que utiliza todas as amostras disponíveis como amostra de treinamento e teste, sendo capaz de alcançar resultados mais precisos que outros métodos de validação cruzada.

Ainda de acordo com Schreiber (2017), em uma base de dados com 100 registros, por exemplo, é definido $k = 10$ (10-*fold)*, então a base de dados será dividida

em 10 subconjuntos. Assim, um subconjunto é utilizado para a validação e os outros 9 subconjuntos são utilizados para treinamento. Então, o processo de validação cruzada é realizado 10 vezes ($k = 10$), até que todos os $k$ subconjuntos tenham sido utilizados na validação. Esse processo é repetitivo pois tem o objetivo de aumentar a confiabilidade da precisão do classificador.

*b)* Método *hold-out*

Segundo Schreiber (2017), o método *hold-out* é semelhante ao método $k$*-fold*, porém a base de dados é dividida em apenas dois conjuntos, onde uma das partes é usada para treinamento e a outra para teste. Diferente do método $k$*-fold,* esse processo é realizado apenas uma vez.

### 2.4.2. Avaliação de Modelo

Segundo os estudos de Wang *et al.* (2020), para avaliar o modelo do algoritmo são utilizadas algumas métricas, como acurácia, sensibilidade e precisão, que podem ser calculadas pela Matriz de Confusão.

Para construir a matriz de confusão, nas colunas são identificadas as classes reais (output) e nas linhas são identificadas as classes preditas (classificação prevista pelo algoritmo), ou seja, essa matriz é um teste para comparar qual classificação o algoritmo previu com a classificação real da instância.

**Quadro 1 – Matriz de Confusão**

| | **Real Classe SIM** | **Real Classe NÃO** |
|---|---|---|
| **Predita Classe SIM** | VP | FP |
| **Predita Classe Não** | FN | VN |

Fonte: Adaptado e Traduzido de Wang *et al.* (2020)

Onde:

**VP** – Verdadeiro Positivo
**FP** – Falso Positivo
**FN** – Falso Negativo
**VN** – Verdadeiro Negativo

a) Acurácia

Acurácia é uma métrica de avaliação que tenta explicar quão bom é o modelo de classificação utilizado para “acertar” as classificações. Ela pode ser calculada a partir do número de acertos que se obtém em relação a uma matriz de confusão pela equação (2):

$$\text{Acurácia} = \frac{VP+VN}{VP+FP+VN+FN} \quad (2)$$

b) Sensibilidade ou *Recall*

Essa outra métrica consegue avaliar quão bom o modelo do algoritmo é para classificar a classe de interesse em relação às instâncias que foram classificadas como “SIM”, como é demonstrado na fórmula a seguir:

$$\text{Sensibilidade} = \frac{VP}{VP+FN} \quad (3)$$

Ou seja, de todas as instâncias que são classificadas como “SIM”, quantas o modelo conseguiu classificar corretamente.

c) Precisão

A precisão é parecida com a sensibilidade, porém, ela avalia de todas as instâncias que o modelo classificou como “SIM”, quantas delas realmente são dessa classe, como é mostrado na fórmula abaixo:

$$\text{Precisão} = \frac{VP}{VP+FP} \quad (4)$$

## 2.5. Evasão Universitária

A evasão no ensino superior é um problema internacional que afeta as instituições de ensino e o sistema educacional do país. Ainda assim, de acordo com Silva Filho *et al.* (2007), no Brasil, a grande maioria da IES não possui nenhum programa de combate à evasão.

O Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia realizou coleta de dados e análises sobre a evasão média nos cursos de Bacharelado Presencial no Brasil entre os anos de 2000 e 2005, com base nos dados do INEP, e encontrou uma média anual de 22% nas IES do Brasil. (SILVA FILHO *et al.*, 2007)

A Comissão Especial de Estudos realizou um estudo sobre a Evasão nas Universidades Públicas Brasileiras (1996) intitulado “pioneiro e inovador de indiscutível relevância para o Sistema de Ensino Superior do país” por ter uma abrangência nacional e também por adotar uma metodologia com capacidade de dar uniformidade aos processos de coleta e tratamento de dados.

A pesquisa buscava identificar as principais causas do fenômeno da evasão, podendo ser caracterizadas em fatores internos e externos. Os fatores internos seriam questões relacionadas à estrutura e dinâmica de cada curso, e os fatores externos, seriam as questões relacionadas a variáveis econômicas, sociais, culturais e até mesmo individuais que podem interferir na vida acadêmica dos estudantes. Os fatores individuais seriam problemas relacionados à habilidade de estudo do aluno, personalidade, dificuldade em adaptação à vida universitária, consequências de uma má formação escolar anterior e até mesmo dificuldades decorrentes de uma escolha precoce da profissão. Outra questão é quando a universidade oferece dupla opção de entrada, ou seja, uma segunda opção para o aluno caso ele não consiga uma vaga na primeira escolha. Pois o aluno se matricula no curso escolhido para segunda opção, mas permanece no curso só até o momento em que conseguir uma vaga no curso de sua primeira opção.

## 2.6. Trabalhos Correlatos

Ainda que a Mineração de Dados Educacionais seja uma área de estudo relativamente nova, no Brasil existem estudos realizados por autores que buscaram encontrar os fatores que poderiam contribuir para o alto índice de evasão das universidades brasileiras, ou ainda, o perfil dos estudantes que teriam uma tendência a desistir dos cursos, por meio da aplicação de técnicas de Mineração de Dados. O Quadro 2 mostra alguns estudos na área de Mineração de Dados Educacionais.

**Quadro 2 – Trabalhos Correlatos**

| Autores | Objetivo | *Dataset* | Técnicas |
|---|---|---|---|
| Brito *et al.* (2019) | Identificar padrões; classificar o perfil mais propenso à evasão; identificar os possíveis fatores que contribuem para o crescimento da evasão. | Faixa etária; Residência; Quantidades de matrículas por semestre; Quantidades de reprovações nas disciplinas básicas; Tempo de permanência no curso até a evasão. | *K-means* e Árvore de decisão. |
| Lane e Alcântara (2018) | Gerar modelo de algoritmo capaz de identficar os estudantes que apresentam risco de evasão a partir do seu primeiro ano no curso de graduação. | Faixa etária; Gênero; Escola de Origem; Bolsista; Estado de origem; Média Enem; Intervalo de tempo (conclusão do Ensino Médio e ingresso na Faculdade); Coeficiente de Rendimento; Situação acadêmica. | Árvore de decisão. |
| Manhães *et al.* (2011) | Identificação precoce dos alunos com risco de evasão. | Coeficiente de Rendimento acumulado no período; Nota e Situação (aprovado, reprovado por nota, reprovado por falta) nas principais disciplinas do primeiro período. | Aprendizado de regras; Tabela de decisão; Árvore de decisão; Modelos lineares de regressão logística; Modelo de rede neural artificial; Modelo probabilístico; Classificador probabilístico simples baseado na aplicação do teorema de Bayes. |
| Paz e Cazella (2017) | Identificar perfil de alunos com potencial de evasão. | Currículo do aluno; Campus; Incentivo; Data de Nascimento; Semestre atual; Município; Evasão. | Árvore de decisão. |
| Xu *et al.* (2019) | Identificar diferenças nos comportamentos online entre grupos de desempenho e prever o desempenho dos alunos a partir dos dados de uso da Internet. | Duração *online*; Duração *offline*; Volume de Internet; Frequência de conexão (móvel e PC). | Árvore de decisão, Máquina de Vetor de Suporte e k-NN. |

| | | | |
|---|---|---|---|
| Coussement *et al.* (2020) | Melhorar previsões de evasão escolar através da comparação do algoritmo LLM com outros oito algoritmos e especificar os impactos da demografia dos alunos; características da sala de aula; e variáveis de envolvimento acadêmico, cognitivo e comportamental no abandono escolar. | 122 variáveis do aluno como variáveis independentes nos modelos de previsão da evasão, classificados em dados: demográficos, características da sala de aula, envolvimento cognitivo, envolvimento acadêmco e envolvimento comportamental. | *Logit Leaf Model;* Árvore de decisão; Regressão logística; Redes neurais; Máquina de vetores de suporte; *Random forest; Boosting tree; Hidden Markov models; Naive Bayes; Bayesian network.* |
| Oztekin (2016) | Prever a taxa de graduação dos alunos e identificar a importância dos fatores que impulsionam a graduação. | Principais variáveis: Média de Nota (Outono), Situação de Moradia, Ensino Médio, Média de Nota (Primavera). | Árvore de decisão, Redes neurais artifi-ciais e Máquinas de vetores de suporte. |
| Lykourentzo *et al.* (2009) | Propor um método de previsão de evasão para cursos *e-learning* a partir de técnicas de aprendizado de máquina. | Gênero; Residência; Experiência de trabalho; Alfabetização da língua inglesa; Nota de teste de múltipla escolha; Nota do projeto; Data de envio do projeto (dias a contar do prazo da seção); Atividade da seção. | Redes neurais *Feed-forward, Probabilistic ensemble simplified fuzzy ARTMAP* e Máquinas de vetores de suporte. |

Fonte: Elaborada pela autora (2021)

Brito et al. (2019) realizaram estudos a respeito da evasão no curso de Sistema de Informação da Universidade Federal do Rio Grande do Norte - UFRN utilizando técnicas de Mineração de Dados. Os autores utilizaram a ferramenta WEKA para executar o algoritmo de agrupamento $k$-means e o algoritmo de classificação J48 para aplicação da técnica da árvore de decisão. Com estes experimentos foi possível encontrar evidências para algumas questões, como os fatores que possivelmente mais colaboram para a evasão do aluno, que no caso seria reprovar nas quatro disciplinas base do curso, não fazer parte de nenhum projeto de pesquisa, estar incluso na faixa de idade entre 26 anos ou mais, exceder os 8 semestres do curso e/ou reprovar nas disciplinas de Fundamentos e Algoritmos.

Lanes e Alcântara (2018) realizaram sua pesquisa a partir dos dados coletados de 12 cursos de graduação da Universidade Federal do Rio Grande – FURG, fazendo o uso de técnicas de Mineração de Dados Educacionais para gerar um modelo de

classificação capaz de identificar o perfil dos estudantes que apresentam risco de evasão já no primeiro ano do curso de graduação. No estudo foi utilizado o algoritmo J48 da ferramenta WEKA para realizar a tarefa de classificação e gerar a árvore de decisão, mostrando que os potenciais alunos evasores podem ser identificados pelo algoritmo com uma acurácia de 90,7%.

Manhães et al. (2011) realizaram um estudo sobre alunos em risco de abandono no curso de Engenharia Civil na Escola Politécnica da Universidade Federal do Rio de Janeiro – UFRJ. Neste estudo, os autores trabalharam com 10 algoritmos diferentes de Mineração de Dados (aprendizado de regras, tabela de decisão, árvore de decisão, modelos lineares de regressão logística, modelo de rede neural artificial, modelo probabilístico e classificador probabilístico simples baseado na aplicação do teorema de Bayes) e as aplicaram em dois ambientes diferentes da ferramenta WEKA, o Weka Experiment Environment (WEE), que realiza comparações entre o desempenho de vários algoritmos de mineração de dados, e o Weka Explorer (WE), que permite a seleção e execução de um algoritmo classificador por vez. Após realizado os experimentos, os resultados mostraram que a partir das primeiras notas semestrais dos alunos já é possível fazer a previsão de alunos com risco de evasão.

Paz e Cazella (2017) realizaram um estudo da Universidade Comunitária do Rio Grande do Sul, através do processo de Mineração de Dados, aplicando a tarefa de classificação com a técnica de árvore de decisão utilizando o WEKA. Os experimentos realizados com essas técnicas tiveram uma acurácia acima de 90% e foi possível evidenciar que os incentivos (bolsas) e currículos dos alunos estão diretamente relacionados à tendência de evasão, uma vez que essas evasões tendem a ocorrer nos semestres iniciais dos cursos com alunos sem esses incentivos.

Xing Xu et al. (2019) realizaram um estudo na China sobre a relação entre o comportamento dos alunos na Internet e seus desempenhos acadêmicos. Além disso, utilizaram modelos de aprendizado de máquina supervisionado para conseguir descobrir se é possível utilizar os dados de uso na Internet para prever o desempenho acadêmico desses alunos. Os autores encontraram evidências de que os alunos com maior tráfego de Internet tinham menores desempenhos acadêmicos, pois a usam para lazer (jogos, chats, vídeos etc.). Nos experimentos preditivos foram utilizados três algoritmos para prever o desempenho acadêmico dos alunos e os resultados mostraram uma acurácia na predição de até 72,75%, mostrando que os dados de uso

da Internet dos alunos são eficazes para prever o desempenho acadêmico dos mesmos.

Coussement et al. (2020) utilizaram os dados de um provedor global de aprendizagem online baseado em assinatura para comparar o algoritmo LLM em relação a outros oito algoritmos, uma vez que, segundos os autores, esses outros modelos não conseguirem explicar a heterogeneidade entre os alunos, ou seja, são modelos de classificação pouco compreensíveis. Já o LLM é uma abordagem híbrida onde ele detecta segmentos nos dados usando os nós-folha de uma árvore de decisão e, em seguida, aplica-se regressão logística a cada segmento, e, de acordo com os resultados, dentro do contexto proposto, é possível prever o abandono escolar com precisão usando cinco tipos de variáveis: demografia; características da sala de aula; e formas cognitivas, acadêmicas e comportamentais de engajamento.

Oztekin (2016), realizou um estudo em uma universidade pública de quatro anos nos EUA, para desenvolver uma abordagem híbrida de análise de dados, onde foram utilizadas técnicas de classificação para prever a conclusão do curso dos alunos de graduação e uma análise de sensibilidade para identificar a importância de cada fator que poderia impulsionar a graduação. Os resultados mostraram que os principais fatores para a previsão das taxas de graduação são as médias das notas do outono, a situação de moradia do aluno e qual escola o aluno frequentou.

Lykourentzou et al. (2009) realizaram um estudo com o objetivo de propor um modelo de previsão de evasão de alunos para cursos online, a partir de três técnicas de aprendizado de máquina e três esquemas de decisão, além dos dados detalhados dos alunos. A pesquisa foi realizada em dois cursos online do Laboratório de Tecnologia Multimídia da Universidade Técnica Nacional de Atenas, na Grécia. De acordo com os autores, os resultados das previsões foram melhores do que a literatura. Além disso, o método proposto também foi avaliado em relação a sua acurácia, sensibilidade e precisão, que também obteve resultados elevados. Com isso, foi possível observar que o esquema foi acurado tanto na identificação correta das evasões quanto na prevenção de erros de classificação completos.

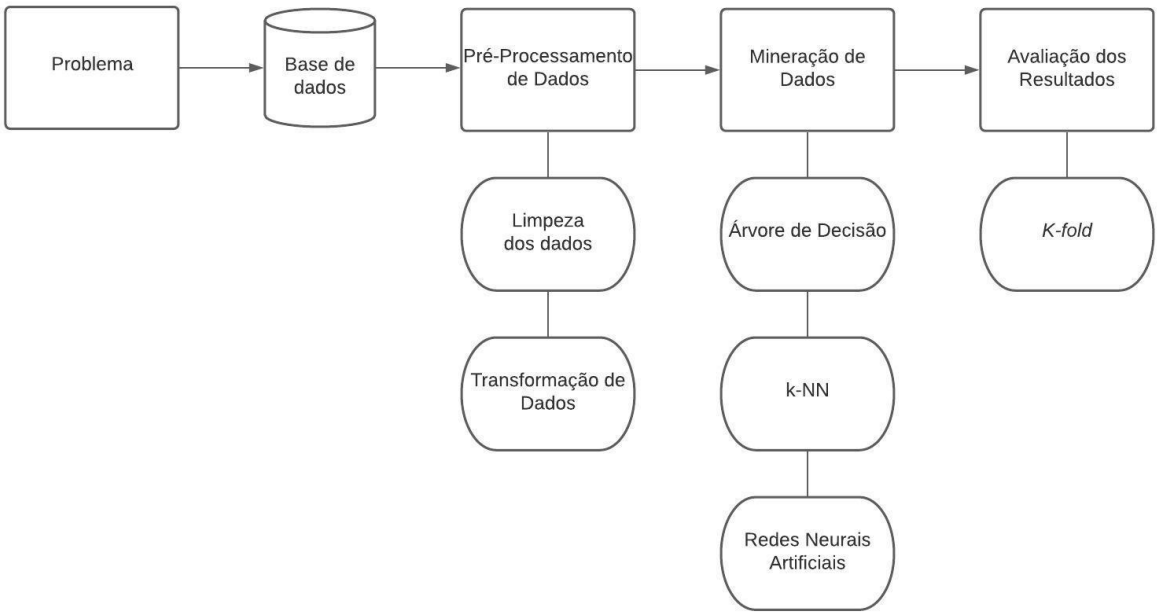
# 3. METODOLOGIA

Neste capítulo descreve-se a base de dados utilizada no estudo e os passos que foram adotados para realizar o pré-processamento dos mesmos, e ainda, o procedimento para aplicação das técnicas de Mineração de Dados.

## 3.1. Etapas do Processo

A Figura 5 mostra o fluxograma com as etapas de todo o processo realizado desde a identificação do problema até a aplicação das técnicas e avaliação dos resultados.

**Figura 5 – Fluxograma das Etapas do Processo**




Fonte: Elaborada pela autora (2021)

Primeiramente é identificado o problema do estudo, a evasão universitária, em seguida é obtido o conjunto de dados, a partir dele é realizado o pré-processamento dos dados, onde é realizada a limpeza e transformação de parâmetros e instâncias. Após essa etapa, é realizada então a mineração de dados, onde são aplicadas as técnicas de mineração: árvore de decisão, RNA e k-NN. E a última etapa é a avaliação dos resultados através do método *k-fold*.

## 3.2. Descrição do Conjunto de Dados

Para a realização desse estudo foi necessário um conjunto de dados mais completo, onde fosse possível analisar diversas variáveis em um grande número de dados. Por isso, a base de dados deste trabalho foi coletada do próprio sistema da instituição, que armazena diversas informações sobre os alunos, desde sua situação acadêmica até informações socioeconômicas. Essas informações são obtidas a partir de um questionário socioeconômico que o aluno responde ao acessar o “Portal do Aluno”. Para ter acesso a esses dados foi necessário obter autorização da Diretoria de Graduação da universidade, esclarecendo a necessidade dos dados para o estudo e garantindo a preservação da identidade dos alunos.

A coleta dos dados foi realizada no ano de 2019 e o conjunto de dados a princípio foi obtido com informações de 691 alunos do curso de Engenharia de Produção de uma universidade pública do Norte do Paraná, e a partir dele, foi possível dar início a fase de pré-processamento de dados.

## 3.3. Pré-Processamento

O conjunto de dados contava com 15 parâmetros: “Sexo”, “Data de Nascimento”, “Idade”, “Nascimento”, “Estado”, “Coeficiente”, “Período do Aluno", “Escore do Processo Seletivo”, “Ano Ingresso”, “Semestre de Ingresso”, “Forma de Ingresso”, “Cotista”, “Estudou em Escola Pública”, “Grupo Étnico” e “Situação”.

Primeiramente, na variável “Situação” foi identificado classificações de alunos com situações acadêmicas que fugiam do objetivo da pesquisa ou que teriam pouca interferência no resultado, como por exemplo, alunos com a seguintes situação: “Desistente (sem cursar)”, “Trancado”, “Transferência”, “Mudou de Curso”, “Falecido”, “Afastado para estudo no exterior” e “Formado”, restando apenas dados de alunos em situação “Regular” ou “Desistente”.

Outro ponto foi a transformação da variável em relação ao local de nascimento do aluno, onde foi necessário binarizar esse atributo, ou seja, foi criado um outro parâmetro que diz se o local de nascimento do aluno é “0” (para os alunos que moram na região onde a universidade está localizada) ou “1” (fora da região). Isso porque esse parâmetro irá dizer se o fato de o aluno sair de casa para estudar e morar longe de sua cidade natal pode influenciar na sua evasão. Com a transformação desse parâmetro, também foi possível excluir a variável “Estado”.

Os atributos “Sexo”, “Cotista”, “Estudou em Escola Pública”, “Forma de Ingresso” e “Grupo Étnico” também foram binarizados. Visto que o parâmetro “Cotista” tinha muitas subcategorias, ele passou a ser classificado em “0” (ampla concorrência) ou “1” (cotista). A variável “Estudou em Escola Pública” passou a ser classificada em “0” (particular) e “1” (pública) e o parâmetro “Sexo” também passou a ser classificado em “0” (feminino) e “1” (masculino).

Na variável “Forma de Ingresso” havia poucos dados de alunos que ingressaram por “Aproveitamento de curso”, “Complementação da lista de espera do SISU” e “Transferência ex-ofício", por isso esses dados foram excluídos. Também foram excluídos os dados dos alunos de “Transferência”, pois foi identificado que os mesmos estavam com dados do “Escore do Processo Seletivo” faltando, deixando para análise apenas dados dos alunos que ingressaram pelo “SISU” ou “Reopção de curso”. Assim, o atributo “Forma de Ingresso” passou a ser classificado em “0” (Reopção) ou “1” (SISU).

Dentro da variável do “Grupo Étnico” havia 6 classificações: “Branca”, “Amarela”, “Preta”, “Pardo”, “Indígena” e “Não declarado”. Em primeiro lugar, foi excluída a classe “Indígenas”, pois havia dados de apenas um aluno para essa classe. Em segundo lugar, para realizar a binarização, foi necessário modificar essa variável, em vez de ser apenas um atributo (“Grupo Étnico”), cada classificação do grupo passou a ser um atributo, e então, cada uma delas foi classificada em “0” (não) e “1” (sim).

Foram excluídas algumas variáveis irrelevantes para o estudo, como “ID”, que era apenas a identificação do aluno na lista do conjunto de dados, “Data de Nascimento”, uma vez que já havia a variável “idade” e o “Semestre de Ingresso”, visto que já havia o ano em que o aluno ingressou na universidade.

Por último, também foi necessário excluir dados que estavam incompletos, como por exemplo, instâncias faltando dados do parâmetro “Escore do Processo Seletivo” e “Coeficiente” (valor “0”).

O Quadro 3 a seguir mostra quais foram os atributos avaliados e suas codificações:

**Quadro 3 – Atributos selecionados e suas codificações**

| Atributos Brutos | Formato original | Transformado para | Utilizado para o estudo? |
|---|---|---|---|
| ID | Categórico (nominal) | - | Não |
| Sexo | Categórico (nominal) | Binário (binarização) | Sim |
| Data de Nascimento | Data | Numérico (novo atributo “idade”) | Sim |
| Nascimento (local) | Categórico (nominal) | Binário (novo atributo “região)” | Sim |
| Estado | Categórico (nominal) | - | Não |
| Coeficiente | Numérico (contínuo) | - | Sim |
| Período do Aluno | Categórico (ordinal) | - | Sim |
| Escore Processo Seletivo | Numérico (contínuo) | - | Sim |
| Ano Ingresso | Numérico (discreto) | - | Sim |
| Semestre Ingresso | Numérico (discreto) | - | Não |
| Forma de Ingresso | Categórico (nominal) | Binário (binarização) | Sim |
| Cotista | Categórico (nominal) | Binário (novo atributo) | Sim |
| Estudou em | Categórico | Binário | Sim |

| | | | |
|---|---|---|---|
| Escola Pública | (nominal) | (binarização) | |
| Grupo Étnico | Categórico (nominal) | - | Sim |
| Situação | Categórico (nominal) | Binário (binarização) | Sim |

Fonte: Elaborado pela autora (2021)

Assim, a base de dados para este estudo após o tratamento do conjunto de dados contou com 16 parâmetros e com os dados de 478 alunos, sendo 340 “Desistentes” e 138 “Regulares”, onde foi possível avaliar os atributos de gênero, idade, local de nascimento, coeficiente de rendimento, escore do processo seletivo, ano de ingresso, período atual, forma de ingresso, se o aluno é cotista, se o aluno estudou em escola pública, grupo étnico e situação acadêmica do aluno.

## 3.4. WEKA

Para realizar todo o processamento de dados foi utilizado a ferramenta *WEKA* (*Waikato Environment for Knowledge Analysis)*, um *software* desenvolvido pela Universidade de Waikato, Nova Zelândia (disponível em: https://www.cs.waikato.ac.nz/ml/weka/). Essa ferramenta fornece algoritmos diferentes para se trabalhar com mineração de dados e aprendizado de máquina (SINGHAL; JENA, 2013).

De acordo com Witten e Frank (2005), na ferramenta *WEKA*, é possível realizar um pré-processamento do conjunto de dados, aplicar um método de aprendizagem e analisar o desempenho do classificador sem precisar fazer a programação do mesmo. O sistema trabalha com métodos para todos os problemas de mineração de dados padrão: regressão, classificação, *clustering*, regra de associação e seleção de atributos.

Também é possível aplicar os métodos de validação do algoritmo, bem como obter as métricas de avaliação, para analisar a confiabilidade do algoritmo utilizado no treinamento dos dados.

# 4. RESULTADOS E DISCUSSÃO

Para encontrar as informações relevantes para a evasão dos alunos foram realizados alguns experimentos. Utilizando a ferramenta *WEKA*, foram utilizados os algoritmos J48 (Árvore de Decisão - AD), IBk (k-NN) e *MultilayerPerceptron* (Redes Neurais Artificiais - RNA), a fim de comparar seus desempenhos e encontrar o mais adequado para atender o objetivo da pesquisa.

## 4.1. Experimento 1

No experimento 1 foram realizadas análises com as técnicas de AD, k-NN e RNA utilizando todas as variáveis e instâncias definidas na etapa de pré-processamento de dados. A Tabela 1 mostra os resultados obtidos no Experimento 1:

**Tabela 1 – Resultados Experimento 1**

| Técnica | Acurácia | Precisão | Recall |
|---|---|---|---|
| AD | 88,5% | 82,7% | 76,1% |
| k-NN (k=1) | 81,4% | 68,7% | 65,2% |
| k-NN (k=3) | 82,8% | 73,0% | 64,5% |
| k-NN (k=5) | 83,2% | 76,4% | 60,9% |
| k-NN (k=15) | 83,7% | 84,1% | 53,6% |
| RNA (tx. ap. = 0,3) | 86,4% | 80,7% | 69,6% |
| RNA (tx. ap. = 0,1) | 87,4% | 79,5% | 76,1% |

Fonte: Elaborada pela autora (2021)

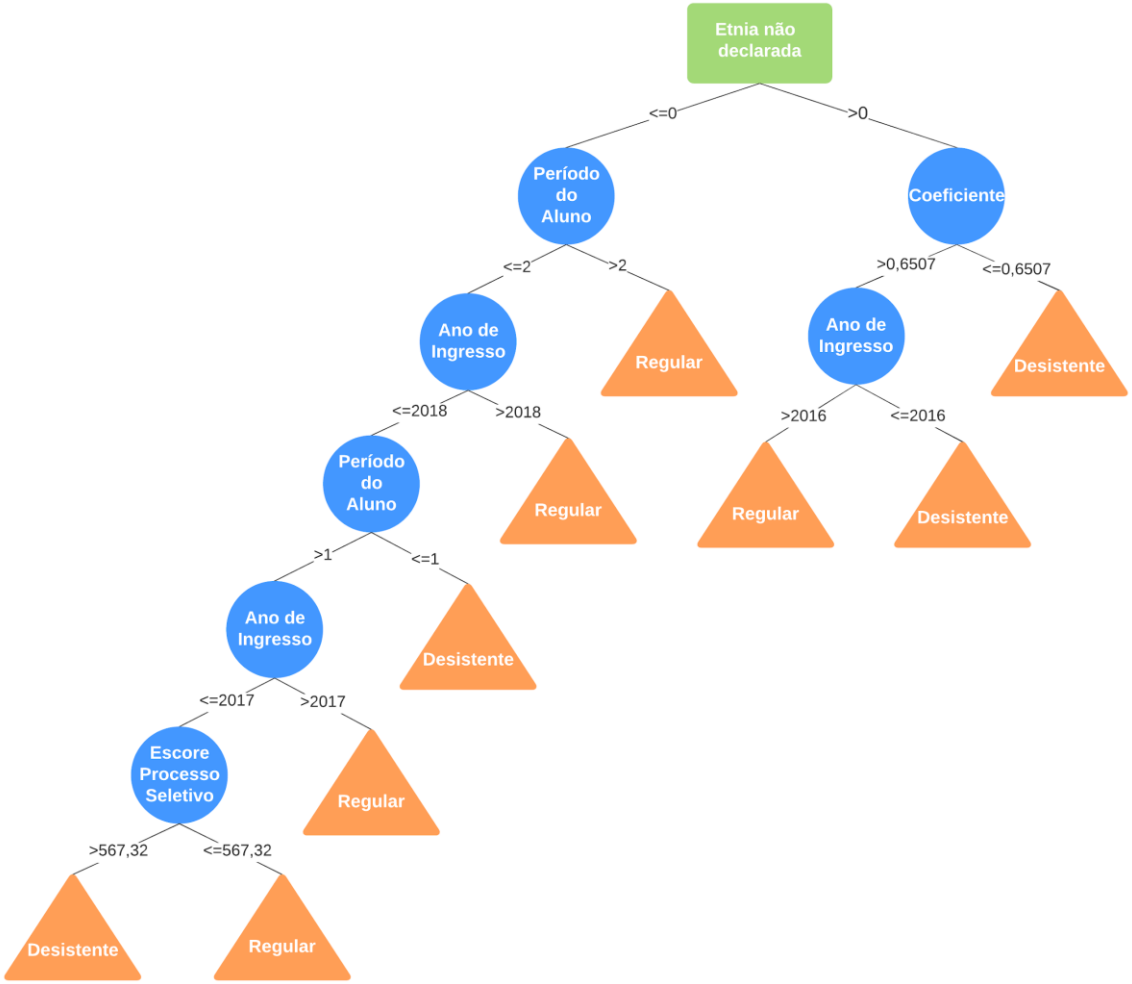
Todas as análises obtiveram uma boa acurácia, porém, é importante salientar que a análise das métricas de avaliação deve ser feita levando em consideração o desempenho das três métricas e nunca de apenas uma. Como é o caso da técnica do k-NN (k=1) que obteve uma acurácia de 81,4%, mas teve uma precisão de 68,7%, ou seja, o algoritmo foi bom para acertar as classificações realizadas, porém de todas as classificações que o modelo classificou como “desistente”, apenas 68,7% eram realmente desta classe. A técnicas de k-NN (k=15) obteve uma acurácia ainda maior, 83,7%, precisão de 84,1%, mas teve uma sensibilidade (*Recall*) de 53,6%, o que quer dizer que de todos os “desistentes” o algoritmo conseguiu classificar somente 53,6%.

As técnicas de k-NN e RNA foram testados com valores diferentes em seus parâmetros para comparar seus resultados e identificar quais têm os melhores

desempenhos entre si. A técnica que obteve o melhor resultado neste primeiro experimento foi AD, com melhor desempenho em acurácia, 88,5%, precisão, 82,7% e *recall*, 76,1%.

A Figura 6 mostra a árvore gerada pelo algoritmo e os atributos mais relevantes para encontrar esse resultado, uma vez que o próprio algoritmo seleciona as variáveis mais importantes durante a análise.

**Figura 6 – Árvore gerada pelo algoritmo J48**




Fonte: Adaptado do WEKA (2021)

Este tipo de técnica de seleção de variáveis que existe dentro do próprio algoritmo de *machine learning* é chamado de técnica de seleção embutida (*embedded*). Esse método de seleção de atributos é diferente do método de filtro, onde o processo de seleção de variáveis ocorre antes da aplicação do algoritmo, utilizando as características do conjunto de dados para selecionar os atributos relevantes, e do método *wrapper*, onde são gerados diversos subconjuntos de

atributos e a cada iteração se utiliza o algoritmo para analisar esse subconjunto selecionado (John *et al.*, 1994; Blum e Langley, 1996).

Essa seleção permite identificar quais são os possíveis atributos que podem ter mais influência na evasão dos alunos, porém, é importante entender quais destes estão sendo relevantes para o algoritmo. Alguns podem ter sido selecionados pelo algoritmo por ter influência maior, de fato, na separação das classes, mas outros podem ter sido selecionados apenas por uma lógica de programação do próprio algoritmo e podem não ter nenhum significado prático na evasão acadêmica, ou seja, não se pode obter algum direcionamento para o combate à evasão.

## 4.2. Experimento 2

Com a seleção de variáveis realizada pela AD no Experimento 1, foi possível realizar outro experimento, testando os algoritmos com os atributos reduzidos, ou seja, aqueles selecionados para pertencer à árvore gerada por AD. As análises foram realizadas com as técnicas de k-NN e RNA com as mesmas variações de parâmetros e com as variáveis: “Etnia=não declarada”, “Período do Aluno”, “Coeficiente”, “Ano de Ingresso” e “Escore do Processo Seletivo”. Os resultados encontrados estão resumidos na Tabela 2.

**Tabela 2 – Resultados Experimento 2**

| Técnica | Acurácia | Precisão | *Recall* |
|---|---|---|---|
| k-NN (k=1) | 90,4% | 84,8% | 81,2% |
| k-NN (k=3) | 89,1% | 85,2% | 75,4% |
| k-NN (k=5) | 88,7% | 87,5% | 71,0% |
| k-NN (k=15) | 86,8% | 85,7% | 65,2% |
| RNA (tx. ap. = 0,3) | 90,6% | 85,5% | 81,2% |
| RNA (tx. ap. = 0,1) | 92.3% | 89,8% | 82,6% |

Fonte: Elaborada pela autora (2021)

Foi possível observar que tanto a técnica k-NN quanto RNA obtiveram desempenhos melhores com a seleção de atributos, mostrando que ambos possivelmente atingem um sobreajuste (*overfitting*) ao utilizar todos os atributos, isso quer dizer que, segundo Castanheira (2008), o modelo obteve um ótimo desempenho com o conjunto de dados para treino, mas acaba aprendendo apenas sobre esse

conjunto de dados (treino), e não consegue ter um bom desempenho ao classificar o conjunto de dados para teste (novas instâncias). Este fato é importante para destacar o quão generalizável o modelo se torna.

Neste cenário, com os atributos reduzidos, o algoritmo que apresentou melhor desempenho foi o de RNA (taxa de aprendizagem = 0,1). Caso fosse necessário selecionar um modelo para a classificação de possíveis alunos para a evasão, a melhor sugestão seria o de RNA (taxa de aprendizagem = 0,1) apenas com os atributos selecionados no Experimento 2, visto que foi o modelo que obteve os melhores resultados, com Acurácia de 92,3%, Precisão de 89,8% e *Recall* de 82,6%.

## 4.3. Comparativos do k-NN e RNA

As Tabelas 3 a 5 trazem comparações entre os desempenhos dos algoritmos k-NN e RNA no primeiro experimento, quando foi realizado com todos os atributos do conjunto dados, com os resultados do segundo experimento, depois de realizada a seleção de atributos. A Tabela 3 mostra a diferença das técnicas aplicadas com o conjunto de atributos reduzidos em relação ao conjunto completo (em termos de acurácia).

**Tabela 3 – Comparativo Resultados para Acurácia**

| | | **Atributos Reduzidos** | | | | | |
|---|---|---|---|---|---|---|---|
| | | k-NN (k=1) | k-NN (k=3) | k-NN (k=5) | k-NN (k=15) | RNA (tx. ap.=0,3) | RNA (tx. ap.=0,1) |
| **Todos Atributos** | k-NN (k=1) | 9,0% | 7,7% | 7,3% | 5,4% | 9,2% | 10,9% |
| | k-NN (k=3) | 7,5% | 6,3% | 5,9% | 4,0% | 7,7% | 9,4% |
| | k-NN (k=5) | 7,2% | 5,9% | 5,5% | 3,6% | 7,4% | 9,1% |
| | k-NN (k=15) | 6,7% | 5,4% | 5,0% | 3,1% | 6,9% | 8,6% |
| | RNA (tx. ap.=0,3) | 4,0% | 2,7% | 2,3% | 0,4% | 4,2% | 5,9% |
| | RNA (tx. ap.=0,1) | 2,9% | 1,7% | 1,3% | -0,6% | 3,1% | 4,8% |

Fonte: Elaborado pela autora (2021)

A Tabela 4 traz o comparativo em relação a acurácia das técnicas e mostra uma melhoria significativa entre RNA, principalmente com taxa de aprendizagem de 0,1. Com relação às outras técnicas, essa diferença é ainda maior quando comparado com a técnica k-NN (k=1), sendo de 10,9%. Um ponto importante a ser destacado é a diferença com valor negativo entre a técnica de k-NN (k=15) e RNA (taxa de aprendizagem de 0,1), isso quer dizer que, mesmo com a seleção de atributos, a

técnica k-NN (k=15) continuou tendo um desempenho pior que o de RNA com taxa de aprendizagem de 0,1.

**Tabela 4 – Comparativo Resultados para Precisão**

| | | Atributos Reduzidos | | | | | |
|---|---|---|---|---|---|---|---|
| | | k-NN (k=1) | k-NN (k=3) | k-NN (k=5) | k-NN (k=15) | RNA (tx. ap.=0,3) | RNA (tx. ap.=0,1) |
| **Todos Atributos** | k-NN (k=1) | 16,1% | 16,5% | 18,8% | 17,0% | 16,8% | 21,1% |
| | k-NN (k=3) | 11,8% | 12,2% | 14,5% | 12,7% | 12,5% | 16,8% |
| | k-NN (k=5) | 8,4% | 8,8% | 11,1% | 9,3% | 9,1% | 13,4% |
| | k-NN (k=15) | 0,7% | 1,1% | 3,4% | 1,6% | 1,4% | 5,7% |
| | RNA (tx. ap.=0,3) | 4,1% | 4,5% | 6,8% | 5,0% | 4,8% | 9,1% |
| | RNA (tx. ap.=0,1) | 5,3% | 5,7% | 8,0% | 6,2% | 6,0% | 10,3% |

Fonte: Elaborado pela autora (2021)

Em relação aos resultados da precisão, não houve nenhum valor negativo, o que significa que houve melhoria em relação a todos as técnicas após a seleção de atributos. Ainda assim, da mesma forma que a acurácia, a técnica que obteve o melhor progresso em relação a precisão, foi as RNAs com taxa de aprendizagem de 0,1, sendo de 21,1% maior que a da técnica do k-NN (k=1).

**Tabela 5 – Comparativo Resultados para *Recall***

| | | Atributos Reduzidos | | | | | |
|---|---|---|---|---|---|---|---|
| | | k-NN (k=1) | k-NN (k=3) | k-NN (k=5) | k-NN (k=15) | RNA (tx. ap.=0,3) | RNA (tx. ap.=0,1) |
| **Todos Atributos** | k-NN (k=1) | 16,0% | 10,2% | 5,8% | 0,0% | 16,0% | 17,4% |
| | k-NN (k=3) | 16,7% | 10,9% | 6,5% | 0,7% | 16,7% | 18,1% |
| | k-NN (k=5) | 20,3% | 14,5% | 10,1% | 4,3% | 20,3% | 21,7% |
| | k-NN (k=15) | 27,6% | 21,8% | 17,4% | 11,6% | 27,6% | 29,0% |
| | RNA (tx. ap.=0,3) | 11,6% | 5,8% | 1,4% | -4,4% | 11,6% | 13,0% |
| | RNA (tx. ap.=0,1) | 5,1% | -0,7% | -5,1% | -10,9% | 5,1% | 6,5% |

Fonte: Elaborado pela autora (2021)

Em relação ao *recall* (sensibilidade), mesmo com a seleção de atributos, a técnica k-NN, apesar de ter uma melhora comparando quando utilizado todos os atributos, a maioria não conseguiu superar a técnica de RNA com taxa de aprendizagem de 0,1 com todos os atributos. Neste comparativo, a técnica de RNA com taxa de aprendizagem de 0,1 também obteve a maior taxa de melhoria, mas agora em relação a técnica k-NN (k=15).

# 5. CONCLUSÕES

Os modelos de aprendizado de máquina aplicados no trabalho tiveram como objetivo extrair informações relevantes sobre a evasão na universidade, buscando realizar uma predição da classificação desses alunos quanto a evasão, gerando também hipóteses de quais os fatores mais relevantes para este fenômeno.

Entre as técnicas aplicadas para realizar os experimentos, a que obteve o melhor desempenho, a princípio, a técnica da Árvore de Decisão (AD), e, após a seleção de atributos realizada pelo próprio algoritmo da técnica de AD, a técnica de Redes Neurais Artificiais (RNA), com taxa de aprendizado de 0,1, obteve um desempenho superior a todos os outros modelos aplicados, demonstrando ser o algoritmo mais adequado para realizar a predição da evasão dos alunos.

Os modelos aplicados no estudo também foram necessários para gerar discussões a respeito dos parâmetros analisados. A partir da seleção de atributos realizada pela técnica AD foi possível identificar os atributos mais relevantes para a classificação dos alunos e que podem ter uma maior influência na evasão dos mesmos. Os atributos selecionados pelo algoritmo foram o “Etnia=não declarada”, “Período do Aluno”, “Coeficiente”, “Ano de Ingresso” e “Escore do Processo Seletivo”.

É importante deixar claro que o atributo “Etnia=não declarada” foi selecionado nesse caso por fazer sentido para o algoritmo e não para aplicação na situação real do estudo. O que difere dos outros atributos selecionados, que de fato, têm relação com o desempenho do aluno no curso, mostrando a possibilidade da evasão dos alunos estar relacionada principalmente com esses fatores, ou seja, quanto pior o desempenho do aluno, maior a probabilidade dele evadir.

Portanto, a predição da classificação dos alunos e o levantamento dos principais atributos é importante para gerar discussões a respeito de medidas e políticas que a universidade pode tomar a fim de diminuir a evasão dos discentes. É fundamental que a universidade tome medidas para o acompanhamento de alunos com baixo desempenho e programas de auxílio, não somente a nível acadêmico, mas também a nível psicológico.

Espera-se que este trabalho possa contribuir com pesquisas futuras, como comparação com resultados de outros métodos, como Regressão Logística, Máquina de Vetor de Suporte (*Support Vector Machines* - SVM) e *Random Forest*. Também

sugere-se como aprimoramentos para pesquisas futuras a implementação de atributos mais específicos em relação às questões sociais e econômicas.

Alguns estudos trazem relevância para as questões socioeconômicas e demográficas, porém, os experimentos realizados com a aplicação dos métodos de mineração de dados não tiveram os mesmos. A pesquisa socioeconômica realizada no portal do aluno tem questões mais específicas nesse aspecto social, econômico e demográfico, mas não foram incluídas no conjunto de dados disponibilizado pelo sistema da instituição para o estudo. Portanto, em pesquisas futuras, sugere-se a inclusão de parâmetros que abordem essas questões para tentar melhorar o modelo.

## REFERÊNCIAS


BAKER, R.; ISOTANI, S; CARVALHO, A. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, vol. 19, no. 2, 2011.

BLUM, A.; LANGLEY, P. **Selection of relevant features and examples in machine learning**. Artificial Intelligence, ed. 97, 245-271, 1997.

BRITO, I. et al. **Uso de Mineração de Dados Educacionais para a classificação e identificação de perfis de Evasão de graduandos em Sistemas de Informação da UFRN**. Anais do VIII Congresso Brasileiro de Informática na Educação, p. 10, 2019.

CAMILO, C.; SILVA, J. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Relatório Técnico – Instituto de Informática, Universidade Federal de Goiás, 2009.

CARNIEL, Mauricio. **Aplicação de algoritmos de Mineração de Dados para Identificação de Fatores que Influenciam a Evasão de Alunos do Curso de Ciência da Computação da UNIVALI**. Trabalho Técnico-científico de Conclusão de Curso (Graduação em Ciência da Computação) – Centro de Ciências Tecnológicas da Terra e do Mar, Universidade do Vale do Itajaí, Itajaí, 2013.

CASTANHEIRA, Luciana. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. Belo Horizonte, 2008. Dissertação apresentada como requisito parcial para conclusão do Mestrado Profissional em Computação Aplicada.

COUSSEMENT, K. et al. **Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model**. Decision Support Systems, 135, 2020, 113325.

COVER, T. M.; HART, P. E. **Nearest Neighbor Pattern Classification**. IEEE Transactions on Information Theory, vol. 13, no. 1, p. 7, 1967.

CUNNINGHAM, P.; DELANY, S. **Featureless Similarity**. Relatório Técnico UCD-CSI-2007-1, 2007.

DEKKER, G.; PECHENIZKIY M.; VLEESHOUWERS, J. **Predicting Students Drop Out: A Case Study**. Educational Data Mining, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. **From Data Mining to Knowledge Discovery: An Overview**. AI Magazine, vol. 17, no. 3. AAAI, Providence, 1996.

FERNEDA, E. **Redes neurais e sua aplicação em sistemas de recuperação de informação**. Ciências da Informação, Brasília, v. 35, n. 1, p. 25-30, 2006.

GARCIA, Simone. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde**. Porto Alegre, 2003. Dissertação apresentada como requisito parcial para conclusão do Mestrado Profissional em Computação Aplicada.

HOED, Raphael. **Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação**. Brasília, 2016. Dissertação apresentada como requisito parcial para conclusão do Mestrado Profissional em Computação Aplicada.

JOHN, G., KOHAVI, R., PFLEGER, K. **Irrelevant features and the subset selection problem**. Machine Learning Proceedings 1994. Morgan Kaufmann, 1994. 121-129.

LANES, M.; ALCANTARA, C. **Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados.** XXIX Simpósio Brasileiro de Informática na Educação – SBIE, vol. 29, no. 1. Porto Alegre, 2018.

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

LYKOURENTZOU, I. et al. **Dropout prediction in e-learning courses through the combination of machine learning techniques**. Computers & Education, v. 53, n. 3, p. 950-965, 2009.

MANHÃES, L. et al. **Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados**. XXII Simpósio Brasileiro de Informática na Educação – SBIE, vol. 1, no. 1. Aracaju, 2012.

MCCUE, Colleen. **Data mining and predictive analysis: Intelligence gathering and crime analysis**. Butterworth-Heinemann, 2014.

OLIVEIRA, Ângelo. **Redes Neurais Artificiais Aplicadas na Detecção, Classificação e Localização de Defeitos em Linhas de Transmissão**. Juiz de Fora, 2005. Dissertação apresentada como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Elétrica do programa de Pós-Graduação em Engenharia Elétrica da Universidade federal de Juiz de Fora.

OZTEKIN, A. **A hybrid data analytic approach to predict college graduation status and its determinative factors**. Industrial Management & Data Systems, vol. 116, no. 8, pp. 1678-1699. Massachusetts, 2016.

PAZ, F.; CAZELLA, S. **Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária**. Anais dos Workshops do Congresso Brasileiro de Informática na Educação, vol. 6, no. 1. 2017.

SCHREIBER, J. et al. **Técnicas de Validação de Dados para Sistemas Inteligentes: Uma Abordagem do Software SDBayes**. XVII Colóquio Internacional de Gestão Universitária, Mar del Plata, p. 18. 2017.

SILVA FILHO, R. et al. **A evasão no ensino superior brasileiro.** Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia. Cadernos de pesquisa, v. 37, n. 132, p. 641-659, 2007.

SILVA FILHO, R.; LOBO, M. **Como a mudança na metodologia do INEP altera o cálculo da evasão**. Instituto Lobo, Mogi das Cruzes, 2012.

WANG, Y. et al. **A Comparative Assessment of Credit Risk Model Based on Machine Learning – a case study of bank loan data**. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019). Procedia Computer Science, vol. 174, p. 141-149, 2020.

WITTEN, Ian; FRANK, Eibe. **Data Mining – Practical Machine Learning Tools and Techniques**. Morgan Kaufmann Publishers, 2005.

XU, X. et al. **Prediction of academic performance associated with internet usage behaviors using machine learning algorithms**. Computers in Human Behavior 98, p. 166-173, 2019.