**INSTITUTO FEDERAL GOIANO – CAMPUS CERES**
**BACHARELADO EM SISTEMAS DE INFORMAÇÃO**
**ANA LUIZA GOMES DE SOUZA**

# CLASSIFICAÇÃO E IDENTIFICAÇÃO DE PERFIS DE EVASÃO DO ENSINO MÉDIO INTEGRADO DO CAMPUS CERES DO IF GOIANO

**CERES – GO**
**2021**

**ANA LUIZA GOMES DE SOUZA**

# CLASSIFICAÇÃO E IDENTIFICAÇÃO DE PERFIS DE EVASÃO DO ENSINO MÉDIO INTEGRADO DO CAMPUS CERES DO IF GOIANO

Trabalho de curso apresentado ao curso de Bacharelado em Sistemas de Informação do Instituto Federal Goiano – Campus Ceres, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação, sob orientação do Prof. Me. Adriano Honorato Braga.

**CERES – GO**

**2021**

Sistema desenvolvido pelo ICMC/USP
Dados Internacionais de Catalogação na Publicação (CIP)
**Sistema Integrado de Bibliotecas - Instituto Federal Goiano**

S729c

Souza, Ana Luiza Gomes de
CLASSIFICAÇÃO E IDENTIFICAÇÃO DE PERFIS DE EVASÃO DO ENSINO MÉDIO INTEGRADO DO CAMPUS CERES DO IF GOIANO / Ana Luiza Gomes de Souza; orientador Adriano Honorato Braga. -- Ceres, 2021.
27 p.

Monografia (Graduação em Bacharelado em Sistemas de Informação) -- Instituto Federal Goiano, Campus Ceres, 2021.

1. IF Goiano. 2. Técnico Integrado. 3. Evasão. 4. Mineração de dados. 5. Predição. I. Braga, Adriano Honorato, orient. II. Título.

Responsável: Johnathan Pereira Alves Diniz - Bibliotecário-Documentalista CRB-1 n°2376

INSTITUTO FEDERAL
Goiano

**Repositório Institucional do IF Goiano - RIIF Goiano**
**Sistema Integrado de Bibliotecas**

## TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano, a disponibilizar gratuitamente o documento no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

**Identificação da Produção Técnico-Científica**

| | |
|---|---|
| [ ] Tese | [ ] Artigo Científico |
| [ ] Dissertação | [ ] Capítulo de Livro |
| [ ] Monografia – Especialização | [ ] Livro |
| [X] TCC - Graduação | [ ] Trabalho Apresentado em Evento |

[ ] Produto Técnico e Educacional - Tipo: _____

Nome Completo do Autor: Ana Luiza Gomes de Souza
Matrícula: 2017103202030244
Título do Trabalho: Classificação e identificação de perfis de evasão do ensino médio integrado do Campus Ceres do IF Goiano.

**Restrições de Acesso ao Documento**

Documento confidencial: [**X**] Não [ ] Sim, justifique: _____

_____

Informe a data que poderá ser disponibilizado no RIIF Goiano: __/__/__
O documento está sujeito a registro de patente? [ ] Sim [**X**] Não
O documento pode vir a ser publicado como livro? [ ] Sim [**X**] Não

## DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

1. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
2. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
3. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Ceres, 21/03/2021.

Ana Luiza G. de Souza
_____
Assinatura do Autor e/ou Detentor dos Direitos Autorais

Ciente e de acordo:

Adriano Honorato Braga
_____
Assinatura do(a) orientador(a)

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

# ATA DE DEFESA DE TRABALHO DE CURSO

Ao(s) 02 dia(s) do mês de março do ano de dois mil e vinte e um, realizou-se a defesa de Trabalho de Curso da acadêmica Ana Luiza Gomes de Souza, do Curso de Bacharelado em Sistemas de Informação, matrícula 2017103202030244, cujo título é "Classificação e identificação de perfis de evasão do ensino médio integrado do Campus Ceres do IF Goiano" . A defesa iniciou-se às 19 horas e 05 minutos, finalizando-se às 19 horas e 31 minutos. A banca examinadora considerou o trabalho APROVADO com média 8,7 no trabalho escrito, média 9,3 no trabalho oral, apresentando assim média aritmética final de 9,0 pontos, estando a estudante APTA para fins de conclusão do Trabalho de Curso.

Após atender às considerações da banca e respeitando o prazo disposto em calendário acadêmico, a estudante deverá fazer a submissão da versão corrigida em formato digital (.pdf) no Repositório Institucional do IF Goiano – RIIF, acompanhado do Termo Ciência e Autorização Eletrônico (TCAE), devidamente assinado pelo autora e orientador.

Os integrantes da banca examinadora assinam a presente.

*(Assinado Eletronicamente)*
Adriano Honorato Braga - Orientador

*(Assinado Eletronicamente)*
Membro da banca
Lucas José de Faria

*(Assinado Eletronicamente)*
Membro da banca
Marcel Ferrante Silva

Documento assinado eletronicamente por:

- **Marcel Ferrante Silva, Marcel Ferrante Silva - Professor Avaliador de Banca - Universidade Federal de Goiás (01567601000143)**, em 12/03/2021 15:53:13.
- **Lucas Jose de Faria, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/03/2021 22:06:29.
- **Adriano Honorato Braga, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/03/2021 21:16:04.

Este documento foi emitido pelo SUAP em 02/03/2021. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 244721
Código de Autenticação: efbf5b5f25

Campus Ceres
Rodovia GO-154, Km.03, Zona Rural, None, CERES / GO, CEP 76300-000
(62) 3307-7100

## RESUMO

Apesar de toda a evolução e transformação que a educação sofreu até os dias atuais, o país ainda carece de desenvolvimento, especialmente para dirimir os índices de evasão escolar, que gera consequências na sociedade como um todo. Portanto, objetiva-se a construção de um modelo preditivo a partir de algoritmos de mineração de dados que melhor se adequem às condições e dados advindos dos cursos técnicos integrados ao ensino médio do campus Ceres do IF Goiano. Para tal, utiliza-se a metodologia KDD, para nortear todo o processo. A amostra foi obtida na base de dados do sistema acadêmico e do processo seletivo, ambos fornecidos pela instituição contemplada na pesquisa. O conjunto de dados, após limpeza e transformação dos dados, totalizou 1.478 matrículas com 28 atributos, sendo que 158 dessas matrículas são de evasão. Também foi utilizada a ferramenta Weka e algoritmos que a mesma disponibiliza. Em prol de sugerir o melhor modelo, se fez necessária a comparação dos algoritmos mais utilizados na literatura, são eles: árvore de decisão, SVM, kNN, *naive bayes*, RNA e regressão logística. Após comparação, o algoritmo *Naive Bayes* foi considerado o melhor preditor para o presente conjunto de dados por apresentar uma baixa taxa de erro (16,3%) acerca de alunos evasores e uma boa acurácia (89,6%).



**Palavras-chave:** IF Goiano. Técnico Integrado. Evasão. Mineração de dados. Predição.

## ABSTRACT

Despite all the changes and transformations that education has undergone to the present day, the country still lacks development, especially to resolve school dropout rates, which has consequences for society as a whole. Therefore, the goal here is to build a predictive model based on a data mining algorithm that best fits the conditions and data arising from the technical courses integrated into the high school on the Campus Ceres of the IF Goiano. For this, the KDD methodology is used to guide the entire data mining process. The sample was obtained from the database of the academic system and the selection process, both provided by the institution included in the research. The dataset had a total of 1,769 records which, after cleaning and transforming the data, amounts to 1,478 enrollments, from the years 2015 to 2020, and 28 attributes, with 158 of these enrollments being evasion ones. The Weka tool and algorithms that it provides were also used. To suggest the best model, it was necessary to compare the most used algorithms in the literature, which are decision tree, SVM, kNN, Naive Bayes, RNA, and logistic regression. After comparison, the Naive Bayes algorithm was considered the best predictor for the present dataset, as it delivers a low error rate (16.3%) about dropout students and good accuracy (89.6%).



**Keywords:** IF Goiano. Integrated Technician. Dropout. Data mining. Prediction.

**SUMÁRIO**

INTRODUÇÃO .....8
REVISÃO DE LITERATURA .....10
METODOLOGIA.....14
RESULTADOS E DISCUSSÃO.....19
CONSIDERAÇÕES FINAIS .....23
REFERÊNCIAS.....24

## INTRODUÇÃO

A educação no Brasil percorreu um longo caminho de transformação e reestruturação metodológica, educacional, sociais e histórica (OLIVEIRA; CRUZ, 2017). Tais alterações se configuram como melhorias para as instituições de ensino, o que possibilita a escolarização de milhões de estudantes da educação básica (CALIXTO; SEGUNDO; GUSMÃO, 2017). Os Institutos Federais possuem uma parcela de crédito no desenvolvimento da educação. Já que possuem como propósito a contribuição com o desenvolvimento regional e local por meio da oferta de vagas em cursos qualificantes, técnicos e educação superior (ROSINKE *et al.*, 2020).

Apesar disso, as instituições de ensino ainda enfrentam desafios no caminho para o ensino de qualidade, uma delas é a evasão escolar (SILVA; NUNES, 2015). Segundo a PNP (2018), em 2017, 16,3% dos estudantes abandonaram o curso técnico onde realizaram suas matrículas. Já em 2018 (PNP, 2019), esse número chegou a 42,1%. Nesses anos, o valor corrente por matrícula no Instituto Federal Goiano (IF Goiano), independente do curso, foi de R$16.784,97 e R$15.339,12, respectivamente, sendo esse, um valor sem retorno direto à sociedade pelo fato da não conclusão do curso. Nesse sentido, a evasão escolar atinge tanto a ordem social, como financeira do Brasil, se tornando um dos problemas centrais no ensino (BITENCOURT; FERRERO, 2019).

Quando um estudante evade, além da instituição perder sua eficácia em formar cidadãos para o mercado de trabalho, há perda de recursos financeiros (BRITO JUNIOR *et al.*, 2019). A falta de empregabilidade e a baixa remuneração, também se tornam consequências recorrentes da evasão, já que ela está atrelada a não conclusão do ensino médio (FERREIRA; OLIVEIRA, 2020). Como um dos causadores da distorção idade-série, a evasão escolar afeta indiretamente, além do dano ao próprio indivíduo, o crescimento econômico e a desigualdade social (PORTELLA; BUSSMANN; OLIVEIRA, 2017).

No ensino médio, a evasão é ainda mais intensificada, quando comparada a outras etapas do ensino (IBGE, 2018). Em 2016, 6,6% dos alunos em todo território nacional abandonaram o ensino médio (INEP, 2019). Já em 2018, essa taxa foi reduzida para 6,1% (INEP, 2019). Apesar de uma redução nas taxas, o problema da evasão escolar continua bastante significativo (BRASIL, 2020). Quando se entende o

padrão da evasão escolar, as políticas públicas, que interferem diretamente nas escolas e no sistema de ensino, podem focar naqueles estudantes que se encaixam nesse padrão de risco desde início de sua jornada, a fim de reverter os índices (BEZERRA *et al.*, 2016).

Evitar ao máximo as práticas de evasão promove a eficiência das instituições de ensino, que precisam cada vez mais procurar ferramentas poderosas para entender tal fenômeno (BEZERRA, 2019). Uma delas é a predição de dados utilizando a mineração de dados educacionais. Ela busca obter conclusões particulares a partir de estudos e combinações de dados específicos (COSTA *et al.*, 2012). Com a utilização de modelos de predição, torna-se possível a realização de ações para combater abandono escolar nas instituições de ensino (ROVIRA; PUERTAS; IGUAL, 2017).

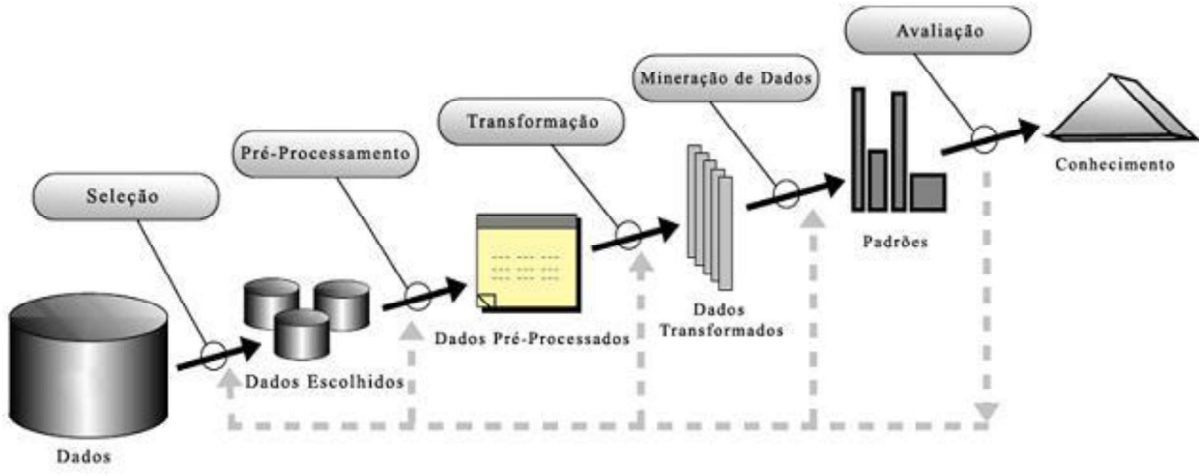
Pelos critérios da PNP (2019), a evasão acontece quando o aluno antes de concluir o curso em que está matriculado o abandona, perdendo o vínculo com a instituição. Inúmeros fatores podem levar alunos a evadirem. Alguns como: tempo na escola, reprovações, localização da escola, condições socioeconômicas, notas baixas, entre outros, podem ser considerados determinantes para evasão (RUMBERGER; LIM, 2008; SILVA FILHO; ARAÚJO, 2017). Tendo isso em vista, este trabalho corrobora com a discussão de alguns dos fatores aqui citados e apresenta os índices dos cursos Técnicos Integrados ao Ensino Médio. Com o auxílio de mineração de dados, criar um modelo de evasão que seja capaz de identificar previamente estes estudantes do Campus Ceres do IF Goiano com perfil de evasão.

## REVISÃO DE LITERATURA

Segundo o *Journal of Educational Data Mining* (2020), a mineração de dados educacionais (MDE, do inglês, *Educational Data Mining*, EDM) é uma disciplina emergente, que objetiva compreender os alunos e os ambientes em que estão inseridos, por meio do desenvolvimento de métodos exploratórios acerca de dados provenientes exclusivamente do âmbito educacional. As técnicas nela utilizadas, na maioria das vezes, são provindas da área de mineração de dados (MD, do inglês, *Data Mining*, DM) (BAKER; ISOTANI; CARVALHO, 2011).

A mineração de dados "consiste na aplicação de análise de dados e algoritmos de descobrimento que produzem uma enumeração de padrões (ou modelos) particular sobre os dados" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 30). Essa, por sua vez, pode ser considerada, juntamente ao pré-processamento dos dados, como uma das principais etapas de um segmento mais amplo conhecido como descoberta de conhecimento em bases de dados (do inglês, *Knowledge Discovery in Databases*, KDD) (COSTA *et al.*, 2012). Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 30) também definem KDD como um "processo não trivial de identificar padrões de dados válidos, novos, potencialmente úteis e, em última análise, compreensíveis", cujas etapas são mostradas na Figura 1.

**Figura 1 - Etapas da descoberta de conhecimento em banco de dados.**




Fonte: https://arquivo.devmedia.com.br/REVISTAS/sql/imagens/127/6/1.jpg (2021).

Os objetivos do processo envolvendo mineração de dados podem ser classificados de duas maneiras diferentes, a predição e a descrição. Na descrição,

procura descrever o comportamento padrão de uma determinada amostra de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Já na predição, busca-se o desenvolvimento de um modelo baseado em uma amostra de atributos capaz de prever situações com valores desconhecidos (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Algumas das técnicas mais utilizadas para predição de dados são: árvore de decisão, máquina de vetores de suporte, *k*-ésimo vizinho mais próximo, redes bayesianas, rede neural artificial e regressão logística.

Árvores de decisão são largamente utilizadas para gerar sistemas de classificação baseados em diferentes atributos. De acordo com Song e Lu (2015), esta é uma técnica não paramétrica, que classifica de forma eficiente um grande e complexo conjunto de dados em um modelo de fragmentos similar a ramos, construindo uma árvore invertida. Ao percorrer a árvore, da raiz às folhas seguindo as condições especificadas em cada um dos nós intermediários, se torna possível identificar os atributos que mais influenciam a classificação e aqueles que podem não ter muito efeito na previsão (GULATI, 2015). Os algoritmos mais conhecidos desse método são o C4.5 (QUINLAN, 1993), C5.0 (RULEQUEST, 2019) e o CART (BREIMAN *et al.*, 1984).

Bitencourt e Ferrero (2019) aplicando mineração de dados com os algoritmos de árvore de decisão *Random Forest*, *Decision Trees*, além do algoritmo *XGBoost*, propõem o uso de descoberta de conhecimento em bases de dados para extrair informações de 772 instâncias de alunos de cursos técnicos que realizaram matrícula entre 01/2018 e 07/2019 no Instituto Federal de Santa Catarina. Ao final dos experimentos foi possível recuperar 25% dos potenciais evasores, com precisão de 86%.

Já em um panorama de graduação, Brito Junior *et al.* (2019) procuraram identificar fatores que pudessem influenciar a evasão no curso de Sistemas de Informação da Universidade Federal do Rio Grande do Norte (UFRN). A partir de 196 instâncias de estudantes, executaram os algoritmos *K-Means*, de clusterização, e J48, de árvore de decisão. Após realização de três experimentos, com precisão de aproximadamente 90 e 95%, constataram quatro diferentes perfis de alunos tendenciosos a evasão, são eles: escolares que reprovaram nas quatro disciplinas base do curso; escolares com idade de 26 anos ou mais; escolares que excederam o tempo comum de conclusão do curso e não participam de nenhum projeto; e escolares

que reprovaram na disciplina de algoritmos, depois de reprovarem na disciplina de introdução a informática.

Outro estudo sobre a evasão foi feito por Tasnim, Paul e Sattar (2019), utilizando um repositório de conjuntos de dados denominado UCI *Machine Learning Repository* aplicaram técnicas de máquina de vetores de suporte, redes bayesianas e regressão logística. Após separação do *dataset* em dois conjuntos, A e B, um contendo 395 instâncias e outro com 649 instâncias, conseguiram alcançar uma precisão de 97% no conjunto A e 100% no conjunto B, ambos com o método de mineração de dados de máquina de vetores de suporte.

As bases para a criação da máquina de vetores de suporte (do inglês, *Support Vector Machine*, SVM) foram desenvolvidas por Vapnik (1995) e vem ganhando popularidade nos últimos anos (CRISTIANINI; SHAWE-TAYLOR, 2000). SVM é um algoritmo supervisionado, eficiente na classificação de dados de alta dimensão, cuja técnica é encontrar um hiperplano ideal em um conjunto de dados (CHENG; TAN; JIN, 2010). Esse hiperplano ideal é aquele que maximiza a distância entre exemplos de treinamento (vetores de suporte) de diferentes classes (GUNN, 1998).

Utilizado também na tarefa de classificação em aprendizado supervisionado, a regressão logística é um modelo linear generalizado relacionado a probabilidade de um evento ocorrer dentre um conjunto de eventos, que pode variar entre 0 e 1. A partir da variável dependente binária, o modelo permite avaliar quais atributos são mais relevantes para que ocorra a variável dependente (HOSMER; LEMESHOW, 2000).

Em sua dissertação de mestrado, Bezerra (2019) utilizou a metodologia CRISP-DM para aplicação de mineração de dados. Com algoritmos de árvore de decisão e redes bayesianas, e com dados acadêmicos e socioeconômicos, do sistema acadêmico do Instituto Federal de Rondônia - Campus Ji-Paraná, pôde identificar dois perfis de discentes com chances de evadir. Entre eles, alunos do primeiro ano letivo, dos turnos matutino e vespertino e alunos que utilizam ônibus coletivo como transporte escolar, possuem baixa renda familiar e que recebem auxílio estudantil.

Os algoritmos de aprendizado de máquina de redes bayesianas são classificadores probabilísticos baseados na aplicação do Teorema de *Bayes*, proposto por Thomas Bayes no século XVIII (BAYES, 1763). Apontado como um dos algoritmos mais completo e simples para classificação de dados, ele utiliza de frequências de co-

ocorrência de cada atributo do conjunto de treinamento para calcular as probabilidades associadas a cada classe de saída (SOMBRA, 2018).

Já as redes neurais artificiais (RNAs) são modelos computacionais inspirados na estrutura do cérebro humano e na capacidade de aprendizado e generalização (HAYKIN, 1994). Geralmente usadas para reconhecimento de padrão por meio de aprendizagem supervisionada, elas utilizam a formação de múltiplas camadas de neurônios que interagem usando conexões ponderadas, além das camadas de entrada e saída, quando se trata de múltiplas camadas (PAL; MITRA, 1992).

Em um estudo comparativo de classificadores de árvore de decisão, SVM, regressão logística, RNA e *k*-ésimo vizinho mais próximo, na tarefa de predição de dados, Ramos *et al.* (2018), usaram 18.094 instâncias do ensino à distância (EaD) da graduação de licenciaturas em Pedagogia e Ciências Biológicas da Universidade de Pernambuco (UPE). Com atributos relacionados à interação dos alunos com a plataforma utilizada no EaD e manuseio da ferramenta *RStudio*, conseguiram alcançar 89,6% de acurácia no preditor de *k*-ésimo vizinho mais próximo e uma classificação correta de 71,9% dos estudantes que evadem com algoritmos de regressão logística.

O classificador *k*-ésimo vizinho mais próximo (do inglês, *k-nearest neighbor, kNN*) prevê qual componente do conjunto de treinamento está mais próximo à instância de teste desconhecida, por meio do cálculo de distância entre as instâncias (BARROS; OLIVEIRA, 2017). O parâmetro k define quantos vizinhos serão escolhidos pelo algoritmo, o que irá impactar no desempenho de diagnóstico do mesmo (ZHANG, 2016). Ele foi desenvolvido a partir de um problema estatístico envolvendo estimativas paramétricas desconhecidas ou difíceis de determinar e teve suas principais regras elaboradas por Cover e Hart (1967).

Entretanto, este trabalho se difere dos demais apresentados por ter como público-alvo especificamente estudantes de cursos técnicos integrados ao Ensino Médio, além disso utiliza de um número maior de atributos comparados aos demais trabalhos citados, incluindo até mesmo atributos anteriores ao ingresso do estudante à instituição como: nota do Processo Seletivo, reserva de vagas e escola de origem. Além da informação sobre a cidade do estudante, já que muitos dos Institutos Federais recebem escolares de todo país, que muitas vezes precisam se deslocar diariamente ou mudarem de residência.

## METODOLOGIA

Para traçar um estudo de caso sobre evasão escolar, optou-se primeiramente pela realização de uma busca na literatura acerca da metodologia mais utilizada para análises preditivas utilizando mineração de dados. Com o auxílio da ferramenta StArt[1] (*State of the Art through Systematic Review*) foi executada uma revisão da literatura, considerando as bases científicas ACM *Digital Library*, IEEE *Xplore Library*, SciELO e Google Acadêmico, em 11 de junho de 2020, de acordo com as *strings* de busca mostradas na Tabela 1.

Em razão do objetivo deste estudo estar inteiramente ligado ao descobrimento de conhecimento com padrões de dados de alunos brasileiros, escolheu-se a alteração das *strings* de busca para as bases que fossem capazes de trazer trabalhos feitos com registro colhidos no território nacional. Sendo assim, para a base científica Google Acadêmico, foram retiradas as *strings* em língua estrangeira, refinando os resultados apenas para estudos nacionais.

**Tabela 1 - Quantidade de resultados retornados de acordo com as *strings* de busca utilizadas nas bases científicas.**

| Base | *String* de busca | Quantidade |
|---|---|---|
| ACM *Digital Library* | [[*All*: "*prediction*"] *OR* [*All*: "predição"]] *AND* [[*All*: "*school*"] *OR* [*All*: "*student*"] *OR* [*All*: "estudante"] *OR* [*All*: "aluno"]] *AND* [[*All*: "*dropout*"] *OR* [*All*: "evasão"]] *AND* [[*All*: "*data mining*"] *OR* [*All*: "mineração de dados"]] | 152 |
| IEEE *Xplore Library* | ("*All Metadata*": "*prediction*" *OR* "*All Metadata*": "predição") *AND* ("*All Metadata*": "*student*" *OR* "*All Metadata*": "*school*" *OR* "*All Metadata*": "estudante" *OR* "*All Metadata*": "aluno") *AND* ("*All Metadata*": "*dropout*" *OR* "*All Metadata*": "evasão") *AND* ("*All Metadata*": "*data mining*" *OR* "*All Metadata*": "mineração de dados") | 42 |
| SciELO | (“*prediction*” OR “predição”) AND (“*student*” *OR* “*school*” *OR* “estudante” *OR* “aluno”) *AND* (“*dropout*” *OR* “evasão”) *AND* (“*data mining*” *OR* “mineração de dados”) | 2 |
| Google Acadêmico | (“predição”) AND (“estudante” *OR* “aluno”) *AND* (“evasão”) *AND* (“mineração de dados”) | 118 |

Fonte: Própria (2020).

[1] http://lapes.dc.ufscar.br/

Após a avaliação dos títulos, resumos e texto completo, foram selecionados 42 trabalhos do total de 314 retornados das bases científicas. Ao comparar as técnicas utilizadas neles, pode-se perceber que há uma carência no uso de dados advindos de cursos técnicos integrados ao ensino médio, muito ofertado em Institutos Federais no Brasil. A prevalência foi nas pesquisas em cursos online e cursos de graduação presenciais, principalmente na área de conhecimento de ciências exatas e da terra. Sendo que, a escolha da quantidade de cursos, em sua maioria, foi de apenas um.

Em geral, os trabalhos não analisaram mais do que 2.000 instâncias, o que pode interferir no resultado dos algoritmos de mineração de dados. Dentre as variáveis escolhidas para integrar o conjunto de dados, predominaram as interações da plataforma, para cursos online, socioeconômicas e de rendimento acadêmico. Por sua vez, as técnicas de mineração de dados mais comuns foram as de árvore de decisão, redes bayesianas, regressão logística, redes neurais, kNN e SVM, respectivamente.

O presente estudo adotou a metodologia KDD para nortear a predição dos dados. Este modelo foi escolhido por satisfazer as características de ser representado em uma linguagem de alto nível, retratar com precisão o conteúdo do banco de dados, gerar resultados interessantes e ser eficiente (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1993). Para tanto, essa técnica precisa contemplar as etapas de seleção, pré-processamento, transformação, mineração dos dados e avaliação.

A etapa de seleção dos dados se deu a partir da obtenção da base de dados do sistema acadêmico do IF Goiano e do processo seletivo (PS) do Campus Ceres. A partir disso, foi preciso desempenhar um estudo acerca das bases para entender como chegar até as variáveis desejadas, já que, principalmente, a base do sistema acadêmico, por se tratar de sua larga utilização em toda rede do IF Goiano, possuía dezenas de tabelas e relacionamentos não necessários para a pesquisa. De modo a unificar o conjunto de dados, realizou-se a integração das informações contidas tanto no Processo Seletivo quanto Sistema Acadêmico através da equidade de dados como: nome, ano de ingresso e curso, dos alunos devidamente matriculados.

A base de dados do sistema Acadêmico do IF Goiano foi adquirida diretamente com a instituição, já que se trata de dados sensíveis. Estes foram disponibilizados por meio de exportação em arquivos na linguagem SQL (*Structured Query Language*), única e exclusivamente para fins de pesquisa. Por outro lado, as informações do PS foram obtidas através dos resultados dos mesmos no site oficial da instituição e

precisaram ser convertidos de arquivo PDF (*Portable Document Format*) para CSV (*Character-separated Values*).

A escolha dos atributos que compõem o *dataset* (Figura 2) procurou abranger a maioria dos atributos encontrados na revisão da literatura. Os alunos contemplados na pesquisa foram os do curso técnico integrado ao ensino médio do Campus Ceres, do IF Goiano, que se matricularam entre os anos de 2015 a 2019. A data-base utilizada para obter o *dataset* final foi obtida no dia 01 de outubro de 2020, totalizando 1.769 alunos. Esses dados foram anonimizados, a fim de preservar a identidade dos alunos e quaisquer informações sensíveis dos mesmos.

**Figura 2 - Atributos selecionados para o modelo preditivo.**

| **Atributo** | **Tipo de dado** | **Possíveis valores** |
|---|---|---|
| Ano do último registro | Numérico | Valores de 2015 a 2020. |
| Cidade | Categórico | 90 valores: Ceres, Rialma, entre outros. |
| Coeficiente de rendimento | Numérico | Valores decimais de 0,0 a 10,0. |
| Cotas | Categórico | 10 Valores: AC, RI, RS, entre outros. |
| Curso | Categórico | 3 valores: Agropecuária, Informática, entre outros. |
| Data de nascimento | Data | Valores no formato AAAA-MM-DD. |
| Escola Origem | Categórico | 3 valores: Pública, estadual e privada. |
| Nota das disciplinas | Numérico | Valores decimais de 0,0 a 10,0. |
| Nota de ingresso | Numérico | Valores inteiros de 0 a 60 |
| Quantidade de faltas | Numérico | Valores inteiros. |
| Quantidade de reprovações | Numérico | Valores inteiros. |
| Série | Numérico | Valores série 1 a 3. |
| Sexo | Categórico | 2 valores: F e M. |
| Situação de matrícula | Categórico | 8 Valores: concluído, evasão, entre outros. |

Fonte: Própria (2020).

A fim de facilitar o pré-processamento dos dados, elaborou-se um Diagrama de Entidade-Relacionamento (DER), na ferramenta gratuita *MySQL Workbench*[2] e posteriormente, construído um banco de dados com o Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL 8.0[3], possibilitando a inserção dos registros

[2] https://www.mysql.com/products/workbench/
[3] https://dev.mysql.com/downloads/mysql/

provindos do sistema acadêmico, por meio da linguagem de consulta estruturada, mais conhecida como linguagem SQL. Assim, foi executada a limpeza, remoção de dados repetidos, *outliers* ou fora do escopo da pesquisa. Ademais, foram retirados os registros referentes às disciplinas em curso, já que as mesmas não possuíam notas.

Foram efetuadas transformações nos dados para adequá-los aos algoritmos, como o cálculo da idade do escolar por meio da diferença entre a data de nascimento e a data do último registro do mesmo na base. A cidade do estudante foi usada para verificar se ele residia na mesma cidade da instituição ou morava em outra cidade. Dessa maneira, essa informação foi transformada em valores booleanos: “sim” para quando o estudante reside na mesma cidade da Instituição de Ensino e “não”, quando não reside na mesma cidade.

Já os cursos foram agrupados em 3 modalidades, “AGRO” contendo o curso de Técnico em Agropecuária Integrado ao Ensino Médio, “INFO” abrangendo os cursos de Técnico em Informática Integrado ao Ensino Médio e Técnico em Informática para Internet Integrado ao Ensino Médio e “MA” relativo ao curso de Técnico em Meio Ambiente Integrado ao Ensino Médio.

Quanto às notas do ensino médio, foram calculadas médias aritméticas por disciplina de núcleo comum cursada para cada aluno, independente da série. Já para as disciplinas de nível técnico, a média foi feita por série independente da disciplina, ou seja, cada série cursada possuía a média das notas obtidas nas disciplinas técnicas da mesma. Além disso, à variável principal, situação de matrícula, foi atribuído “sim” para aqueles que evadiram e “não” para aqueles que concluíram, ainda estão matriculados ou se transferiram. Ao final da transformação totalizaram 28 atributos devido ao detalhamento das notas.

Na fase de mineração dos dados, foi realizada a definição dos algoritmos para indução de modelos preditivos de acordo com as técnicas mais utilizadas pela literatura, sendo eles: J48, *Naive Bayes*, *Logistic*, *Multilayer Perceptron*, *IBk* e *LibSVM*. Estes podem ser encontrados na ferramenta Weka[4] (*Waikato Environment for Knowledge Analysis*), desenvolvida pela Universidade de Waikato na Nova Zelândia, contendo bibliotecas de algoritmos e suporte no processo de mineração de dados (HALL *et al.*, 2009).

[4] https://www.cs.waikato.ac.nz/ml/weka/

Observou-se que a amostra estava desbalanceada com apenas 79 evasões em um total de 1.400 matrículas. Um dos motivos para esse fenômeno é a falta de atualização dos registros por parte da instituição gerenciadora do mesmo. Por essa razão, optou-se pela utilização da técnica SMOTE (*Synthetic Minority Oversampling Technique*) (CHAWLA *et al.*, 2002) de balanceamento para equilibrar artificialmente um conjunto de dados, já que a taxa de evasores se revelou proporcionalmente menor que a taxa de concluintes e matriculados, podendo inferir nos resultados.

Cada um dos algoritmos aqui citados para a realização do estudo, traz consigo parâmetros que ajudam a construir e moldar seus resultados. Estes, por sua vez, não possuem um padrão próprio e correto de valores para cada *dataset*. Umas das estratégias usadas para a definição desses parâmetros baseia-se no teste de diferentes combinações de valores até que se chegue no que melhor se encaixa com os dados e gere melhor resultado.

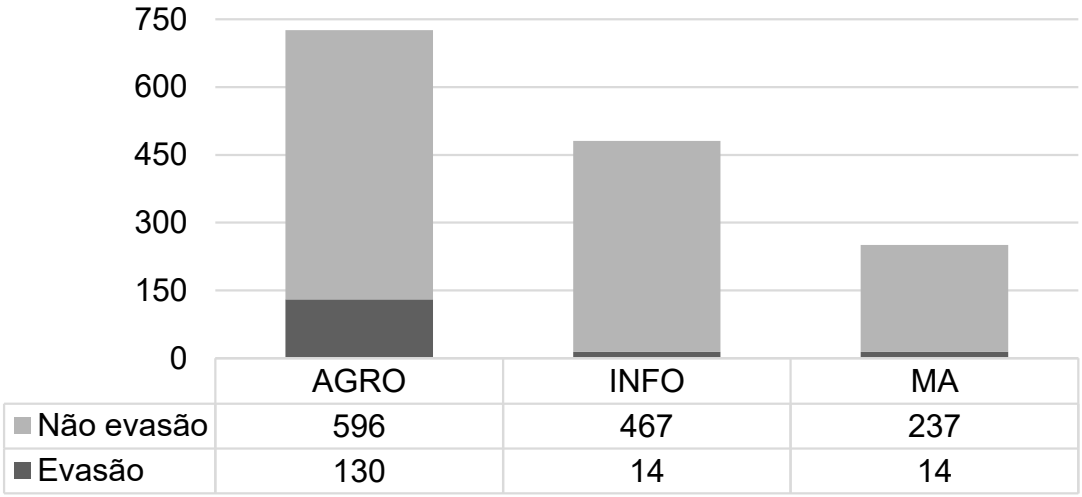
Partindo desse pressuposto, as configurações pré-estabelecidas pelo Weka foram as que obtiveram os melhores resultados para os algoritmos J48, *LibSVM* e *Logistic.* No preditor *Naive Bayes* foi usado o estimador de *kernel* para atributos numéricos em vez de uma distribuição normal, ativando o parâmetro *use Kernel Estimator*. Já no *Multilayer Perceptron* o número de épocas para treinar, *Training Time*, foi aumentado para 750. Por fim, no *IBk*, a função usada para calcular a distância das instâncias foi a *Filtered Distance*, pois esta aplica o filtro antes de chamar a função de distância.

Para cada modelo, foi aplicada a estratégia de validação cruzada dividida em 10 subconjuntos. Posteriormente, foi feita a comparação do desempenho dos algoritmos. Como forma de avaliação, utilizou-se as métricas: acurácia, que diz respeito a taxa de acerto comparado ao desempenho geral; pontuação F1 (F1-*Score*), que é a média harmônica entre precisão e revocação e permite qualificar o modelo de um modo geral; Curva Característica de Operação do Receptor (Curva ROC) que avalia o desempenho do modelo; e taxas de falsos positivos e verdadeiros positivos que identificam a frequência das classes no modelo.

## RESULTADOS E DISCUSSÃO

A base de dados composta pela amostra de 1.478 matrículas, conta com estudantes dos cursos de Agropecuária, Informática, Informática para Internet e Meio Ambiente, todos integrados ao ensino médio ofertados no Instituto Federal Goiano - Campus Ceres. Dessas matrículas, aproximadamente 41% são de estudantes do sexo feminino e 59% do masculino. Em sua maioria, são estudantes vindos de outras cidades que não Ceres, e o ensino fundamental foi concluído em escola estadual. A média de nota no processo seletivo para entrada na instituição, coeficiente de rendimento e nas disciplinas técnicas são de aproximadamente 31, 5,92 e 7,56, respectivamente. O número de estudantes evadidos da amostra foi de 158 e está presente em todos os cursos, assim como mostrado na Figura 3.

**Figura 3 - Gráfico de barras do quantitativo de matrículas por tipo de curso nos anos de 2015 a 2020.**




| | AGRO | INFO | MA |
|---|---|---|---|
| Não evasão | 596 | 467 | 237 |
| Evasão | 130 | 14 | 14 |

Fonte: Própria (2021).

Como métrica de avaliação dos resultados, as matrizes de confusão foram calculadas e estão apresentadas na Figura 4. Esta é composta pela quantidade de erros e acertos de classificação do sistema, de modo a visualizar as amostras confundidas pelo mesmo (SOARES *et al.*, 2020). Para tanto, temos que: Verdadeiro Positivo (VP) é a quantidade de alunos que evadiram e o modelo classificou como evadido, Falso Positivo (FP) é a quantidade de alunos que não evadiram, mas o modelo classificou como evadido, Verdadeiro Negativo (VN) é a quantidade de alunos

que não evadiram e o modelo classificou como não evadido e Falso Negativo (FN) é a quantidade de alunos que evadiram, mas o modelo classificou como não evadido.

A taxa VP atingida por cada preditor, pode ser interpretada como sendo a taxa de recuperação de potenciais estudantes a evadir e a taxa FN como sendo a taxa de perda de possíveis estudantes a evadir. Levando-se em consideração esses aspectos e os valores apresentados na Figura 4 conseguiu-se atingir uma taxa de recuperação de 82,9% com o modelo *Naive Bayes*, que se sobressaiu em relação ao outros modelos. Na contramão da taxa de recuperação, o pior desempenho encontrado foi do classificador *LibSVM*, com perda de predição de 43,7% dos alunos com potencial de evadir.

**Figura 4 - Tabelas das matrizes de confusão para cada classificador.**

| **Modelo** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | VP | FP |
| Não Evadiu | FN | VN |

| **J48** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | 102 | 42 |
| Não Evadiu | 56 | 1278 |

| **IBk** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | 118 | 45 |
| Não Evadiu | 40 | 1275 |

| ***LibSVM*** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | 89 | 13 |
| Não Evadiu | 69 | 1307 |

| ***Logistic*** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | 103 | 55 |
| Não Evadiu | 50 | 1270 |

| ***Multilayer Perceptron*** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | 104 | 40 |
| Não Evadiu | 54 | 1280 |

| ***Naive Bayes*** | **Valor real** | |
|---|---|---|
| **Previsto** | Evadiu | Não Evadiu |
| Evadiu | 131 | 126 |
| Não Evadiu | 27 | 1194 |

Fonte: Própria (2021).

Na Tabela 2 são apresentados os resultados experimentais dos modelos preditivos para o conjunto de dados utilizado. Pode-se observar que a acurácia dos classificadores se manteve acima de 92%, com exceção do algoritmo de redes bayesianas que obteve 89,6%. O modelo *Naive Bayes* foi considerado como modelo de melhor desempenho, pois ainda que tivera a menor acurácia, alcançou a menor porcentagem de FN, ou seja, foi o que menos errou ao tentar predizer alunos evadidos.

O F1-*Score* mostra que o algoritmo *IBk* alcançou um ótimo desempenho. Destaca-se também o classificador *Multilayer Perceptron*, cuja curva ROC foi igual a 0,949. Métricas como a acurácia e a taxa VP foram lideradas pelo *LibSVM*. Vale salientar que mesmo não sendo um dos critérios de avaliação dessa pesquisa, soluções como SVM e rede neurais embora tenham atingido números muito expressivos, levam muito tempo de processamento.

**Tabela 2 - Resultados de classificação das métricas consideradas no estudo.**

| Algoritmo | Acurácia | F1-*Score* | Taxa VP | Taxa FP | Curva ROC |
|---|---|---|---|---|---|
| IBk | 0,942 | **0,943** | 0,942 | 0,230 | 0,867 |
| J48 | 0,934 | 0,932 | 0,934 | 0,320 | 0,923 |
| *LibSVM* | **0,945** | 0,939 | **0,945** | **0,391** | 0,905 |
| *Logistic* | 0,929 | 0,928 | 0,929 | 0,315 | 0,940 |
| *Multilayer Perceptron* | 0,936 | 0,935 | 0,936 | 0,308 | **0,949** |
| *Naive Bayes* | 0,896 | 0,907 | 0,896 | 0,163 | 0,930 |

Fonte: Própria (2021).

Com a obtenção dos resultados pode-se conseguir alguns perfis de evasão da amostra, dentre eles estão, em ordem de significância: estudantes com coeficiente de rendimento menor que 5,6; estudantes no primeiro ano de curso; estudantes com mais de 18 anos; estudantes do curso técnico em agropecuária integrado ao ensino médio; estudantes faltosos; estudantes que obtiveram nota menor ou igual a 25 na prova do processo seletivo; e estudantes advindos de escolas estaduais.

Os resultados aqui encontrados, são relativamente maiores a alguns encontrados na literatura. Maria, Damiani e Pereira (2016) desenvolveram um modelo preditivo com redes bayesianas para classificar a evasão em cursos técnicos não integrados ao ensino médio, obtendo 85,6% de acurácia. Contudo, o modelo não foi escolhido por meio de comparações com outros preditores e, portanto, não foi possível saber o desempenho de outros modelos com o mesmo conjunto de dados utilizado.

Analisando dados de 157.298 discentes de graduação, Do Couto e De Santana (2017) empregaram as mesmas técnicas de mineração aqui citadas, comparando-as e assim escolheu-se o algoritmo *Bayesian Network* com acurácia de 85%. No entanto, apesar de ter usado a forma de ingresso dos alunos, os autores não tinham informação sobre a nota ou reservas de vagas no processo de seleção empregado no ingresso, além de apresentarem indicadores de rendimento acadêmico acumulado própria da instituição de ensino, diferenciando-se do presente trabalho.

Por outro lado, também foram encontrados resultados semelhantes de desempenho, como dos autores Marquez-Vera, Morales e Soto (2013). Eles executaram os algoritmos *JRip*, *NNge*, *OneR*, *Prism*, *Ridor*, *ADTree*, J48, *RandomTree*, *REPTree* e *SimpleCart*, obtendo acurácia maior que 93% em todos eles. Entretanto, a amostra era composta por informações do primeiro ano do ensino médio de 670 alunos, além do conjunto de fatores considerado mais influente: nota menor ou igual a 5,9 em Física e Matemática; nota menor que 4,0 em Humanidades e Leitura e Escrita; nota entre 4,0 e 5,9 em Inglês e Ciências Sociais; idade superior a 15 anos; e nível regular de motivação.

## CONSIDERAÇÕES FINAIS

O ensino técnico integrado no Brasil é amplamente ofertado pela Rede Federal de Educação Profissional, Científica e Tecnológica. Ao mesmo tempo, o mesmo nível de ensino registra altas taxas de evasão e isso acaba se tornando um ponto de alerta. A redução desses índices se torna fundamental para o desenvolvimento de uma nação e para isso é necessária a identificação prévia de alunos com potencial de evasão.

Depois de toda a pesquisa realizada, pode-se observar a importância de predizer estes números a fim de minimizar as altas taxas de evasão, principalmente por meio da utilização da mineração de dados. Este estudo ainda se torna inovador pelo fato de ter trabalhado com cursos técnicos integrados ao ensino médio, pouco explorados até então na literatura. Além de ter atuado com informações de recém-ingresso, ou seja, levando em consideração o conhecimento dos estudantes antes de iniciar o curso, até a saída do estudante, sendo de forma concluída com êxito ou evasão.

Conclui-se que, os modelos preditivos aqui comparados são de total relevância para que juntamente de políticas de acompanhamento da instituição possa ser possível diminuir os números de evasão. O melhor dos casos analisados, o algoritmo *Naive Bayes*, além de auxiliar na predição com boa precisão possui uma pequena taxa de erro sobre a base estudada, ou seja, mais alunos que possivelmente evadirão podem ser englobados nessas medidas de prevenção.

Para trabalhos futuros, sugere-se a adoção de medidas ainda não investigadas como uma forma de aperfeiçoamento das técnicas neste estudo realizadas neste estudo. Podendo assim abordar mais variáveis sobre as atividades realizadas pelos estudantes durante o curso na instituição, além de inserir atributos de fatores externos ao curso, que muitas vezes podem influenciar a decisão dos mesmos sobre sua permanência. Ademais, é possível utilizar a mesma metodologia aqui empregada com os dados dos cursos superiores, comparando assim aos estudos citados na revisão da literatura e ainda ampliar a amostra com a inclusão de dados de outros *campi* do IF Goiano.

## REFERÊNCIAS


BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, 2011. Disponível em: https://doi.org/10.5753/RBIE.2011.19.02.03

BARROS, T. M.; OLIVEIRA, L. A. H. G. de. Modelo Probabilístico para Predição de Desempenho e Evasão de Alunos: um Estudo de Caso no IFRN. **Anais do Workshop de Pesquisa Científica - WPC**, p. 18–23, 2017.

BAYES, T. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. **Philosophical Transactions of the Royal Society of London**, v. 53, p. 370–418, 1763. Disponível em: https://doi.org/10.1098/rstl.1763.0053

BEZERRA, J. H. da S. **Análise da Evasão Escolar do Instituto Federal de Rondônia – Campus Ji-Paraná – Utilizando Técnicas de Mineração de Dados**. 277 f. 2019. - Instituto Politécnico do Porto, 2019. Disponível em: http://hdl.handle.net/10400.22/14490

BEZERRA, C. *et al.* Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes. *In*: **Anais do Simpósio Brasileiro de Informática na Educação**. 2016. p. 1096. Disponível em: https://doi.org/10.5753/cbie.sbie.2016.1096

BITENCOURT, P. B. de; FERRERO, C. Predição de Risco de Evasão de Alunos Usando Métodos de Aprendizado de Máquina em Cursos Técnicos. *In*: **Anais dos Workshops do VIII Congresso Brasileiro de Informática na Educação (CBIE 2019)**. Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2019. p. 149. Disponível em: https://doi.org/10.5753/cbie.wcbie.2019.149

BRASIL. **Evasão Escolar ou Abandono Escolar**. 2020. Disponível em: https://www.gov.br/mdh/pt-br/navegue-por-temas/crianca-e-adolescente/dados-e-indicadores/evasao-escolar-ou-abandono-escolar. Acesso em: 17 set. 2020.

BREIMAN, L. *et al.* **Classification and Regression Trees**. Wadsworth, 1984.

BRITO JUNIOR, I. *et al.* Uso de Mineração de Dados Educacionais para a classificação e identificação de perfis de Evasão de graduandos em Sistemas de Informação. *In*: **Anais dos Workshops do VIII Congresso Brasileiro de Informática na Educação (CBIE 2019)**. Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2019. p. 159. Disponível em: https://doi.org/10.5753/cbie.wcbie.2019.159

CALIXTO, K.; SEGUNDO, C.; GUSMÃO, R. P. de. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. *In*: **Anais [...]**, 2017. p. 1447. Disponível em: https://doi.org/10.5753/cbie.sbie.2017.1447

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. Disponível em: https://doi.org/10.1613/jair.953

CHENG, H.; TAN, P.-N.; JIN, R. Efficient Algorithm for Localized Support Vector Machine. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 4, p. 537–549, 2010. Disponível em: https://doi.org/10.1109/TKDE.2009.116

COSTA, E. *et al.* Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1–29, 2012.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967. Disponível em: https://doi.org/10.1109/TIT.1967.1053964

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. Cambridge University Press, 2000. Disponível em: https://doi.org/10.1017/CBO9780511801389

DO COUTO, D. D. C.; DE SANTANA, Á. L. Mineração de dados educacionais aplicada à identificação de variáveis associadas à evasão e retenção. **Congresso sobre Tecnologias na Educação**, p. 333–344, 2017.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27–34, 1996. Disponível em: https://doi.org/10.1145/240455.240464

FERREIRA, E. C. da S.; OLIVEIRA, N. M. de. EVASÃO ESCOLAR NO ENSINO MÉDIO: causas e consequências. **Scientia Generalis**, v. 1, n. 2, p. 39–48, 2020.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge Discovery in Databases: An Overview. **AI Magazine**, v. 13, n. 3, p. 57–70, 1993. Disponível em: https://doi.org/10.1609/aimag.v13i3.1011

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. 2. ed. Rio de Janeiro – RJ: Elsevier, 2015.

GULATI, H. Predictive analytics using data mining technique. **International Conference on Computing for Sustainable Global Development**, p. 713–716, 2015.

GUNN, S. Support Vector Machines for classification and regression. **Technical Report, Univ. of Southampton**, 1998.

HALL, M. *et al.* The WEKA data mining software. **ACM SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 10–18, 2009. Disponível em: https://doi.org/10.1145/1656274.1656278

HAYKIN, S. **Neural Networks - A Comprehensive Foundation**. 2. ed. Singapore: Prentice Hall, 1994.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. 2. ed. Canada: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2000.

IBGE. **PNAD Contínua 2017: número de jovens que não estudam nem trabalham ou se qualificam cresce 5,9% em um ano**. 2018. Disponível em: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/21253-pnad-continua-2017-numero-de-jovens-que-nao-estudam-nem-trabalham-ou-se-qualificam-cresce-5-9-em-um-ano#:~:text=7%252C2%25).-

,As%2520regi%25C3%25B5es%25. Acesso em: 17 set. 2020.

INEP. **Inep divulga taxas de rendimento escolar; números mostram tendência histórica de melhora**. 2019. Disponível em: http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/inep-divulga-taxas-de-rendimento-escolar-numeros-mostram-tendencia-historica-de-melhora/21206. Acesso em: 17 set. 2020.

Journal of Educational Data Mining. **About the Journal**. 2020. Disponível em: https://jedm.educationaldatamining.org/index.php/JEDM. Acesso em: 24 set. 2020.

MARIA, W.; DAMIANI, J. L.; PEREIRA, M. Rede Bayesiana para previsão de Evasão Escolar. *In*: **Anais dos Workshops do V Congresso Brasileiro de Informática na Educação**. 2016. p. 920. Disponível em: https://doi.org/10.5753/cbie.wcbie.2016.920

MARQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting School Failure and Dropout by Using Data Mining Techniques. **IEEE Revista Iberoamericana de Tecnologias del Aprendizaje**, v. 8, n. 1, p. 7–14, 2013. Disponível em: https://doi.org/10.1109/RITA.2013.2244695

OLIVEIRA, B. C. de; CRUZ, S. P. da S. Verticalização e trabalho docente nos institutos federais: uma construção histórica. **Revista HISTEDBR On-line**, v. 17, n. 2, p. 639, 2017. Disponível em: https://doi.org/10.20396/rho.v17i2.8645865

PAL, S. K.; MITRA, S. Multilayer perceptron, fuzzy sets, and classification. **IEEE Transactions on Neural Networks**, v. 3, n. 5, p. 683–697, 1992. Disponível em: https://doi.org/10.1109/72.159058

PNP. **Plataforma Nilo Peçanha 2018**. 2018. Disponível em: http://plataformanilopecanha.mec.gov.br/2018.html. Acesso em: 19 jun. 2020.

PNP. **Plataforma Nilo Peçanha 2019**. 2019. Disponível em: http://plataformanilopecanha.mec.gov.br/2019.html. Acesso em: 19 jun. 2020.

PORTELLA, A. L.; BUSSMANN, T. B.; OLIVEIRA, A. M. H. de. A relação de fatores individuais, familiares e escolares com a distorção idade-série no ensino público brasileiro. **Nova Economia**, v. 27, n. 3, p. 477–509, 2017. Disponível em: https://doi.org/10.1590/0103-6351/3138

QUINLAN, J. R. **C4.5: programs for machine learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

RAMOS, J. L. C. *et al.* Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. *In*: **Anais [...]**, 2018. p. 1463. Disponível em: https://doi.org/10.5753/cbie.sbie.2018.1463

ROSINKE, J. G. *et al.* A Participação dos Institutos Federais na Interiorização da Educação Superior Presencial no Brasil. **Research, Society and Development**, v. 9, n. 1, p. e06911570, 2020. Disponível em: https://doi.org/10.33448/rsd-v9i1.1570

ROVIRA, S.; PUERTAS, E.; IGUAL, L. Data-driven system to predict academic grades and dropout. **PLOS ONE**, v. 12, n. 2, 2017. Disponível em: https://doi.org/10.1371/journal.pone.0171207

RULEQUEST. **Data Mining Tools See5 and C5.0**. 2019. Disponível em: https://www.rulequest.com/see5-info.html. Acesso em: 24 set. 2020.

RUMBERGER, R.; LIM, S. A. Why Students Drop Out of School: A Review of 25 Years of Research. **California Dropout Research Project, Policy Brief 15, University of California**, 2008.

SILVA FILHO, R. B.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. **Educação Por Escrito**, v. 8, n. 1, p. 35, 2017. Disponível em: https://doi.org/10.15448/2179-8435.2017.1.24527

SILVA, J. L. D. da; NUNES, I. D. Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. **Anais do XXVI Simpósio Brasileiro de Informática na Educação**, p. 1112 – 1121, 2015. Disponível em: https://doi.org/10.5753/cbie.sbie.2015.1112

SOARES, L. C. C. P. *et al.* Aplicação de técnicas de aprendizado de máquina em um contexto acadêmico com foco na identificação dos alunos evadidos e não evadidos. **Revista Humanidades e Inovação**, v. 7, n. 8, p. 224–235, 2020.

SOMBRA, T. R. **Reconhecimento de padrões em rede social científica: aplicação do algoritmo Naive Bayes para classificação de papers no Mendeley**. 198 f. 2018. - Universidade Federal do Rio de Janeiro - UFRJ, 2018.

SONG, Y.-Y.; LU, Y. Decision tree methods: applications for classification and prediction. **Shanghai archives of psychiatry**, v. 27, n. 2, p. 130–135, 2015. Disponível em: https://doi.org/10.11919/j.issn.1002-0829.215044

TASNIM, N.; PAUL, M. K.; SATTAR, A. H. M. S. Identification of Drop Out Students Using Educational Data Mining. *In*: **International Conference on Electrical, Computer and Communication Engineering (ECCE)**. IEEE, 2019. p. 1–5. Disponível em: https://doi.org/10.1109/ECACE.2019.8679385

VAPNIK, V. **The Nature of Statistical Learning Theory**. 1. ed. New York: Springer-Verlag, 1995.

ZHANG, Z. Introduction to machine learning: k-nearest neighbors. **Annals of Translational Medicine**, v. 4, n. 11, 2016. Disponível em: https://doi.org/10.21037/atm.2016.03.37