revista principia Vol. 59, N. 3

REVISTA DE DIVULGAÇÃO CIENTÍFICA E TECNOLÓGICA DO INSTITUTO FEDERAL DE CIÊNCIA, EDUCAÇÃO E TECNOLOGIA DA PARAÍBA

DOI: http://dx.doi.org/10.18265/1517-0306a2021id5488

ARTIGO ORIGINAL

SUBMETIDO 25/02/2021

APROVADO 09/06/2021

PUBLICADO ON-LINE 19/08/2021

PUBLICADO 30/09/2022

EDITORA ASSOCIADA
Crishane Azevedo Freire

Classificação de estudantes com potencial à evasão: aplicando mineração de dados no contexto de cursos técnicos subsequentes do IFPB

Janderson Ferreira Dutra [1]*

João Paulo Lopes de Souza [2]

Damires Yluska de Souza Fernandes [3]

[1] janderson.dutra@ifpb.edu.br. *Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), Campus Cajazeiras, Brasil*

[2] joaopaulopbjp@gmail.com. *Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), Campus Monteiro, Brasil*

[3] damires@ifpb.edu.br. *Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), Campus João Pessoa, Brasil*

*Autor para correspondência.

RESUMO: A mineração de dados educacionais tem sido uma ferramenta muito utilizada para identificar a possibilidade de evasão de estudantes e suas possíveis causas, buscando auxiliar instituições de ensino no acompanhamento, gerenciamento e solução a esse desafio. Este artigo apresenta uma abordagem que aplica mineração de dados educacionais para predizer estudantes de cursos subsequentes do IFPB com potencial de evasão. Para isso, foram coletados dados a partir do sistema acadêmico do IFPB, no contexto do Campus Cajazeiras. Os dados passaram por etapas de preparação, e um conjunto de dados específico foi gerado para o propósito da abordagem proposta. Com a assistência de um especialista de domínio, a abordagem provê a geração de modelos de aprendizado de máquina que classificam a possibilidade de evasão de estudantes, a partir de cinco métodos supervisionados. A avaliação dos métodos de classificação utilizados demonstra que todos os algoritmos apresentaram resultados próximos, a partir do levantamento das métricas obtido. Adicionalmente, o trabalho mostra que a quantidade de períodos cursados é o fator principal para o estudante evadir. O estudo mostra também que a distância geométrica de onde o estudante reside até o campus em questão não é fator relevante para a evasão.

Palavras-chave: aprendizado de máquina supervisionado; evasão de estudantes; mineração de dados educacionais.

Classification of students with dropout potential: applying data mining in the context of subsequent technical courses at the IFPB

ABSTRACT: *The educational data mining area has been widely used as a tool to*

ISSN (on-line): 2447-9187

make predictions on the probability of students evading. Also, it has provided insights on its possible causes to assist educational institutions in monitoring, managing, and solving such challenges. In this sense, this paper presents an approach that applies educational data mining to make predictions on students to identify those who may evade. The students belong to subsequent IFPB courses. With the assistance of a domain specialist, a dataset was produced by using data collected from the IFPB academic system, in the context of the Cajazeiras Campus. The presented approach generates machine learning models that classify the potential for student evasion, based on five supervised methods. The evaluation of the classification methods shows that all the algorithms presented similar results, based on the obtained metrics. Additionally, the work shows that the number of periods taken by students is the main factor for them to escape. This paper also points out that the geometric distance from where the student resides to the campus at hand is not a relevant factor for evasion.

Keywords: *educational data mining; evasion of students; supervised machine learning.*

1 Introdução

A educação tem papel fundamental em levar o conhecimento às diversas classes sociais. Para isso é necessário que as instituições de ensino acompanhem de perto o desempenho acadêmico e as necessidades inerentes dos estudantes.

O problema da evasão escolar ocorre em diversas instituições de ensino, sejam elas privadas ou públicas, e isso tem levado os gestores e a comunidade acadêmica a buscar entender melhor os fatores que levam a esse fenômeno tão recorrente (BRANCO, 2020).

O Ministério da Educação (MEC) (BRASIL, 1996) estabelece que a evasão de curso ocorre quando o estudante sai de forma definitiva do curso de origem sem ter cumprido todos os requisitos obrigatórios para a conclusão. Essa geração corresponde ao ciclo de uma turma no período do ingresso à conclusão. De modo geral, o problema da evasão é pautado por diversas pesquisas que destacam fatores sociais e econômicos como fundamentais para a desistência ou abandono. Vários autores têm trabalhado no entendimento de que fatores levam ao fenômeno da evasão (BÓBÓ *et al*., 2019; CORDEIRO; MUSSA; HORA, 2019; PEREIRA, 2019; QUEIROGA, 2017; SILVA; DIAS; SILVA, 2017).

O problema da evasão de curso atinge não apenas os estudantes no campo acadêmico, mas também compromete todo o contexto social e econômico onde eles estão inseridos. Ou seja, a cada estudante evadido, tem-se uma redução da quantidade de profissionais qualificados no mundo do trabalho, o que piora a economia, bem como aumenta ainda mais a desigualdade social.

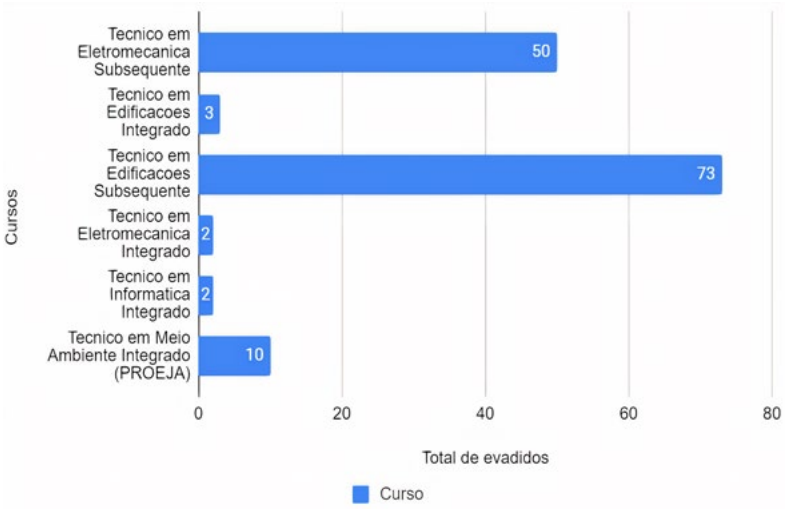
Nos Institutos Federais, a evasão de discentes tem sido um problema preocupante e recorrente. Sobre esse impasse, foi realizada uma pesquisa em um Instituto Federal da região Nordeste, diante de um levantamento referente às causas e consequências. O estudo mostrou que o tema da evasão é alvo de preocupação por parte de profissionais e estudiosos da área, em especial na educação superior. Os dados indicam a necessidade de maior atenção e intervenção pelas instâncias educativas em tentar reduzir esse problema (GUERRA; FERRAZ; MEDEIROS, 2019).

O Instituto Federal da Paraíba (IFPB) oferece cursos de qualificação profissional (FIC), técnicos integrados e subsequentes, bacharelados, licenciaturas, tecnológicos, pós-graduação *lato sensu* e *stricto sensu*. Conforme dados da Plataforma Nilo Peçanha (PNP) (BRASIL, 2018), o IFPB – Campus Cajazeiras apresentou uma taxa de evasão no ano base de 2019 de 13,5% em seus cursos ofertados. Os Cursos Superiores de Tecnologia (CST) apresentaram uma taxa anual de evasão em 2019 de 22,9%. Os Cursos Técnicos Integrado e Subsequente, juntos, apresentaram uma taxa de evasão de 13,0%. Os dados apontados são exemplos de indicadores apresentados pela Plataforma Nilo Peçanha, que é responsável pela consolidação de dados acadêmicos e de gestão acerca da Rede Federal de Educação Profissional, Científica e Tecnológica.

Verificando-se isoladamente, em especial, os Cursos Técnicos Subsequentes apontaram um índice maior, indicando que 24,3% dos estudantes foram evadidos durante o ano de 2019 (BRASIL, 2018).

A Figura 1 mostra uma alarmante diferença entre o quantitativo de estudantes evadidos dos cursos técnicos subsequentes em relação aos integrados no Campus Cajazeiras.

Figura 1 ▶
Número de estudantes evadidos de 2017 a 2019.
Fonte: dados da pesquisa



Juntos, os dois cursos técnicos subsequentes, representam 87,86% do quantitativo de evadidos durante esse período de três anos. Esses números demonstram a dimensão do problema da evasão e a necessidade de uma análise mais aprofundada que possa assistir os gestores na tomada de decisões estratégicas. Para isso, um acompanhamento mais próximo dos estudantes com esse perfil de evasão é imprescindível.

Diante do cenário desafiador associado aos cursos técnicos subsequentes do IFPB, esse estudo se concentra na coleta e uso de dados acadêmicos de estudantes dessa modalidade. Os dados são provenientes do Campus Cajazeiras, da modalidade presencial de técnicos subsequentes considerando o período de 2017 a 2019.

Com base nesses dados, esse trabalho apresenta uma abordagem que aplica modelos de classificação supervisionada para identificar estudantes de cursos subsequentes do Campus Cajazeiras com perfil potencial de evasão. Para isso, duas questões de pesquisa norteiam este trabalho: i) Quais são os principais fatores que influenciam na evasão de estudantes? ii) Qual o melhor modelo de classificação a ser usado para predizer discentes com perfil potencial de evasão?

As análises apresentadas neste artigo foram desenvolvidas com a participação efetiva de um especialista de domínio. Este profissional é oriundo da Diretoria de Desenvolvimento de Ensino (DDE) do Campus Cajazeiras e já vem acompanhando, nos últimos anos, os números de estudantes evadidos no referido *campus*. Logo, o departamento conhece de perto essa realidade. Os resultados apresentados neste trabalho são um recurso a mais para auxiliar essa diretoria nas tomadas de decisões estratégicas no sentido de prevenir a evasão de estudantes que se apresentam com esse perfil. Destarte, também pode auxiliar na redução dos indicadores de evasão apresentados anteriormente, oriundos da Plataforma Nilo Peçanha.

O artigo está organizado da seguinte forma: na seção 2 estão descritos a fundamentação teórica e alguns trabalhos relacionados. A seção 3 explica a metodologia utilizada. Os resultados obtidos são apresentados e discutidos na seção 4. Por fim, a seção 5 trata das considerações finais, algumas limitações encontradas, bem como indica alguns trabalhos futuros.

2 Fundamentação e trabalhos relacionados

A Mineração de Dados Educacionais (EDM, do inglês *Educational Data Mining*) busca desenvolver ou adaptar métodos e algoritmos de mineração de dados existentes, de maneira que possam ser utilizados para compreender melhor os dados em contextos educacionais e gerar modelos úteis de aprendizado (COSTA *et al*., 2012; ROMERO; ROMERO; VENTURA, 2014).

Conjuntos de dados educacionais são constantemente produzidos a partir dos ambientes educacionais onde interagem docentes e discentes. Pesquisas utilizando esses conjuntos de dados têm sido cada vez maiores. Segundo Ramos *et al*. (2020), a EDM auxilia na identificação de variáveis de aprendizagem. Ela é também usada para analisar essas variáveis, a fim de entender o processo de aprendizado apoiado em Tecnologias de Comunicação e Informação (TIC).

Nessa acepção, Rigo *et al*. (2014) afirmam que a ampla difusão de sistemas informatizados, Ambientes Virtuais de Aprendizagem (AVAs), Educação a Distância e aprendizagem híbrida em escolas e universidades favorecem o crescimento do grande volume de dados gerados e armazenados em bases de dados. Os autores ainda acrescentam que esse grande volume de dados tem contribuído para o aumento do interesse na sua utilização junto às técnicas de mineração de dados, haja vista a possibilidade de buscar respostas de perguntas específicas da educação relacionadas aos processos de aprendizagem, desenvolvimento de materiais instrucionais, acompanhamento de estudantes e previsões a partir de informações e padrões comportamentais importantes para determinadas práticas pedagógicas.

Técnicas de EDM têm sido frequentemente utilizadas para (GOTTARDO; KAESTNER; NORONHA, 2014): i) fornecer suporte e mensagens de feedback a professores; ii) recomendações a estudantes; iii) identificação de grupos de estudantes com características comuns; iv) previsão de desempenho ou risco de evasão. Assim, a EDM provê o conhecimento que favorece a melhoria da educação num contexto geral ou específico, por exemplo, avaliando, melhorando ou mesmo mensurando desempenho de estudantes quanto ao seu aprendizado.

Baker, Isotani e Carvalho (2011) descrevem que é possível compreender de forma mais eficaz e adequada os estudantes, buscando entender como eles aprendem e quais fatores influenciam em sua aprendizagem. Os autores ainda exemplificam que é possível

identificar em que situação um tipo de abordagem instrucional proporciona melhores benefícios educacionais aos estudantes, sendo possível predizer o desempenho de um discente com base em alguns atributos comportamentais e usar os resultados apresentados nessa análise para customizar metodologias de ensino para esses indivíduos em particular.

Para viabilizar estudos mais aprofundados e prover recomendações mais assertivas, métodos de aprendizado de máquina (AM) podem ser usados. Goldschmidt, Passos e Bezerra (2015) destacam a classificação como uma das tarefas de mineração de dados mais importantes e mais populares. A tarefa de classificação normalmente faz uso de métodos de aprendizado de máquina supervisionado. Nesta tarefa, os atributos do conjunto de dados são divididos em dois grupos. Um dos grupos contém somente um atributo, que corresponde à variável alvo, ou seja, o atributo para o qual se deve fazer a predição de um valor. O outro grupo contém os atributos a serem utilizados na predição do valor, denominados atributos previsores ou atributos de predição. Os atributos e valores do conjunto de dados usados neste trabalho estão especificados na seção 3.2.

O AM pode ser também não supervisionado. Uma tarefa bastante comum nesse contexto é a de agrupamento, quando busca-se identificar grupos de instâncias conforme similaridades encontradas entre elas. Cada grupo formado pode ser visto como uma classe (GOLDSCHMIDT; PASSOS; BEZERRA, 2015), logo, métodos de agrupamento são também algumas vezes usados para classificação.

Nesse panorama, muitos trabalhos têm sido desenvolvidos na área de EDM para classificar e predizer estudantes propensos à evasão escolar. Santos, Bercht e Wives (2015) desenvolveram uma pesquisa com o objetivo de identificar o estudante desanimado em um ambiente virtual de ensino e aprendizagem, utilizando técnica de árvore de decisão para classificar os estudantes propensos ao desânimo. O estudo foi realizado com estudantes da disciplina Análise de Demonstrações Contábeis ofertada pelo Departamento de Ciências Contábeis e Atuariais da Universidade Federal do Rio Grande do Sul (UFRGS). Os resultados mostraram uma taxa de verdadeiros positivos (*TP rate*) em cerca de 91% de estudantes que são propensos ao desânimo.

De modo semelhante, Machado *et al*. (2018) buscaram investigar aspectos comportamentais de estudantes em ambientes virtuais de aprendizagem por meio de EDM. A pesquisa é aplicada ao recurso de fóruns educacionais, cujo método de AM não supervisionado de agrupamento *k-means* foi utilizado para identificar estudantes com padrões de comportamento comuns nos fóruns. Os autores concluíram que, com a identificação dos grupos de estudantes, a partir da interação nos fóruns de discussões, os professores poderiam obter um suporte educacional para aprofundar o entendimento sobre os aspectos comportamentais. Os resultados favoreceram para que os docentes planejassem melhor as aulas e identificassem as dificuldades de determinados educandos, facilitando e melhorando o processo de aprendizagem a partir de novos métodos de ensino.

Medeiros e Padilha (2018) realizaram um estudo de caso para detectar a evasão de estudantes em uma escola pública da Paraíba. Para isso, utilizaram quatro algoritmos de classificação, através da ferramenta *Weka*: *Part*, *OneR*, *J48* e *RandomTree*. Para esse estudo, os dados foram coletados a partir de um formulário disponibilizado para os estudantes da instituição. Foram utilizadas informações como sexo, localidade de residência, uso de transporte escolar, participação em projeto social, idade, se anteriormente houve abandono em alguma escola, entre outras informações. Sendo três deles baseados em árvores de decisão, os algoritmos de classificação apresentaram resultados diferentes quanto ao motivo de evasão. O algoritmo *Part* registrou que os estudantes do sexo masculino, sem bolsa, evadem devido ao trabalho; resultado semelhante ao obtido pelo *J48*, tendo o trabalho como motivo principal para a evasão.

Já o algoritmo *OneR* destacou a idade como principal atributo utilizado para diferenciar o motivo da evasão. O *Random Tree* gerou regras com o maior nível de detalhe, variando entre idade, trabalho, casamento, bullying ou gravidez, como os principais motivos que levam à evasão dos estudantes.

No trabalho de Medeiros e Padilha (2018) não há explicitação detalhada do algoritmo que obteve um melhor comportamento a partir de cada regra gerada. O artigo não especificou informações claras sobre as métricas utilizadas nos algoritmos. Ademais, os autores sugerem que campanhas contra evasão sejam realizadas pela escola com foco nos estudantes identificados com maior risco à evasão a partir da faixa etária identificada.

Barreto *et al*. (2019) também usaram a técnica de árvore de decisão com o algoritmo *J48* para classificar estudantes de um Instituto Federal. Isso foi realizado a partir de características semelhantes que os estudantes apresentam quando considerados em risco de evasão. Pode-se observar que o comportamento de evasão que mais prevaleceu nos três semestres está ligado a três atributos: idade, curso e sexo. Em relação à idade, isto foi identificado para estudantes que possuem até 21 anos ou acima de 27 anos. Os demais, ou estavam iniciando os estudos no nível superior, ou já estão atuando no mundo do trabalho, investindo em suas carreiras. Sobre o atributo sexo, tem-se uma maior evasão para o masculino em cursos de Licenciatura. E os com maiores chances de evadir são aqueles com idade até 21 anos, quando se trata de curso de Tecnologia. A análise comparativa foi realizada considerando três semestres letivos.

Comparando o presente trabalho a esses relacionados, este investiga os fatores mais preponderantes na evasão de estudantes a partir de dados contidos em um *dataset* real coletado a partir de um sistema acadêmico da Instituição. O trabalho amplia o número e tipos de algoritmos usados e comparados (algoritmos de árvores de decisão, KNN – *K Nearest Neighbors* e SVM – *Support Vector Machine*) para classificar os estudantes. Em tempo, apresenta o grau de importância dos atributos na classificação a fim de compreender os dados com maior potencial em inferir os resultados, como a variável distância. Cabe destacar também nesta presente pesquisa a participação ativa de um especialista de domínio para a compreensão dos dados.

Não foi encontrado na literatura nenhum estudo relacionado à evasão escolar, no contexto do IFPB, que utilize mineração de dados. A análise deste trabalho impulsiona a obtenção de informações que contribuam para o desenvolvimento de políticas públicas dos cursos subsequentes do *campus* em questão, visando a mitigar a evasão escolar.

3 Metodologia

Este trabalho foi desenvolvido com base nas etapas definidas pelo processo CRISP-DM (acrônimo para *Cross Industry Standard Process for Data Mining*) ou Processo Padrão da Indústria Cruzada para Mineração de Dados (IBM CORPORATION, 2017).

O processo CRISP-DM permite organizar fluxos de trabalho, saída e anotações de acordo com as fases de um projeto típico de mineração de dados. É possível produzir relatórios a qualquer momento, durante o projeto, com base nas notas para fluxos e fases do CRISP-DM (CHAPMAN *et al*., 2000).

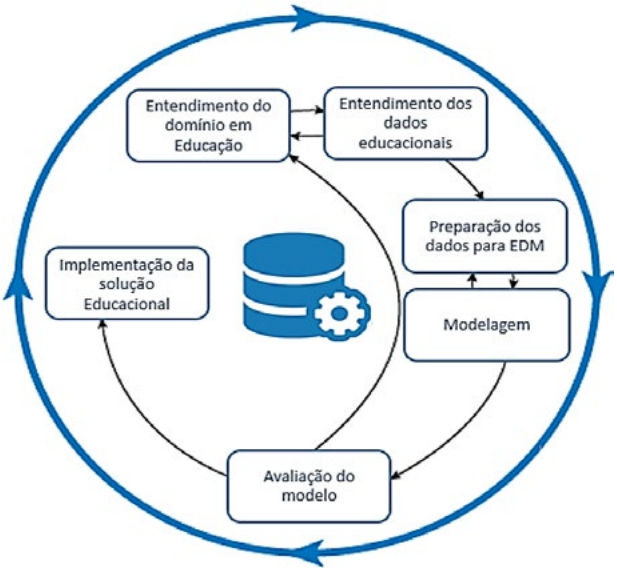
Segundo Ramos *et al*. (2020), com a consolidação do processo CRISP-DM para processos de mineração de dados, os processos de mineração em contextos educacionais, consequentemente, podem usufruir desta metodologia para uma melhor efetivação de projetos de EDM. Ramos *et al*. (2020) sugerem e descrevem detalhadamente uma

extensão ao CRISP-DM aplicado em cenários de mineração de dados educacionais, denominado de CRISP-EDM.

As etapas do CRISP-EDM (Figura 2) são brevemente descritas a seguir (RAMOS *et al*., 2020):

1. Entendimento do domínio em Educação: nesta fase é feito o levantamento das problemáticas educacionais da instituição para as quais se deseja buscar soluções por meio da mineração de dados. Pode ser, por exemplo, um ambiente de *e-learning* de uma universidade, dados do exame nacional de estudantes de ensino médio, dados de sistemas de gestão acadêmica, entre outros. Conhecer bem o domínio da aplicação pode ajudar na definição dos objetivos da mineração e na escolha apropriada de variáveis que serão coletadas para o processo;
2. Entendimento dos dados educacionais: é realizada nessa etapa a análise inicial dos dados de um AVA ou dados coletados em pesquisas como, por exemplo, dados produzidos por um censo escolar. Importante e recomendável que a escolha dos dados seja a partir de alguma teoria educacional já consolidada;
3. Preparação dos dados: envolve a integração de bases de dados, extração, transformação e limpeza das variáveis para a construção do conjunto de dados a ser usado. Como exemplo, pode ser necessária a limpeza dos dados nos logs de acessos dos estudantes ao AVA e transformações de notas dos estudantes (variáveis contínuas) em conceitos (discretas). É importante verificar se há ocorrência significativa de valores fora da normalidade, decorrentes de estudantes muito interativos e valores ausentes, de estudantes com muita, baixa, ou nenhuma interação;
4. Modelagem: nesta etapa são definidas as técnicas de modelagem para geração de modelos de AM. Assim, são definidos os algoritmos a serem aplicados e seus parâmetros. Por exemplo, pode-se definir um modelo de previsão de tendência à evasão de estudantes em Educação a Distância ou, então, uma definição de agrupamento de estudantes de acordo com suas características de interação em um ambiente virtual;
5. Avaliação do modelo: nesta etapa deve-se verificar como os modelos de AM desenvolvidos se comportam com dados reais. É importante validar os modelos de AM, entendendo suas qualidades e características antes de implantá-los em um ambiente real de produção. Um exemplo dessa etapa em um contexto educacional seria avaliar se os modelos preditivos de desempenho dos estudantes atendem às taxas de acurácia, precisão e *recall* compatíveis com outros trabalhos ou com resultados esperados;
6. Implantação do modelo: pode ser feita, por exemplo, a partir do uso de *plug-ins* ou módulos de sistemas que realizam as tarefas de EDM e apresente, por meio de relatórios ou visualizações (e.g., *dashboards*), os resultados da aplicação dos modelos nos dados educacionais.

Figura 2 ▶
Fases do modelo CRISP-EDM.
Fonte: adaptado de Shearer (2000 apud RAMOS et al., 2020)



Com base na metodologia CRISP-EDM, as etapas deste trabalho foram aplicadas e são apresentadas a seguir.

3.1 Entendimento do domínio e dos dados

O Sistema Unificado de Administração Pública (SUAP) é o sistema acadêmico adotado pelo IFPB e contém os dados associados aos históricos escolares dos estudantes e seus dados socioeconômicos.

É possível obter também dados escolares anteriores ao ingresso na instituição como, por exemplo, a instituição frequentada anteriormente ao ingresso no IFPB. Os dados de histórico acadêmico são os mais importantes para o objeto desta pesquisa. Por meio dele, é possível extrair informações quanto a: data de ingresso na instituição, coeficiente de rendimento escolar (CRE), data da conclusão ou data da evasão, curso do qual o estudante participa, e se ele ingressou através de cotas. É possível saber também em quais projetos de pesquisa ou extensão cada estudante participou ou participa.

O conjunto de dados inicial foi gerado em um arquivo .xls pela DDE do campus Cajazeiras. Os dados se referem a registros de todos os estudantes do campus, porém foram selecionados apenas os estudantes dos cursos subsequentes entre os anos de 2017 e 2019. Importante destacar também que foram retirados quaisquer dados sensíveis e que viessem a identificar os envolvidos, estando em acordo com a Lei Geral de Proteção de Dados – Lei nº 13.409/2018.

No subconjunto de estudantes evadidos, não há registros referentes à evasão no ano 2019. Isso se justifica devido ao relatório original não conter essas informações sobre os estudantes ingressantes nesse último ano do período analisado.

Outro ponto importante a mencionar é sobre a não utilização de dados sobre os ingressantes em 2020, visto que a situação causada pela pandemia da COVID-19 culminou na interrupção das atividades presenciais do IFPB a partir do mês de março de 2020. Sua utilização possivelmente geraria um ruído entre os dados coletados durante esse período, devido à possibilidade de terem ocorrido trancamentos ou desistências por essa situação atípica na saúde da população.

3.2 Preparação dos dados

O *dataset* utilizado precisou passar várias vezes pela etapa de pré-processamento para tarefas de limpeza, discretização e redução da dimensionalidade, pois, após a realização dos testes iniciais (preliminares) nos algoritmos, as métricas que são apresentadas na seção 4 não apresentaram um resultado considerado satisfatório. Havia muita instabilidade nos resultados, com valores discrepantes a cada execução dos algoritmos.

Os atributos selecionados foram escolhidos com base na análise do especialista de domínio, que identificou quais seriam os mais importantes para o estudo, sendo eles: curso, distancia, zona, renda, cota, idade, sexo, CRE, ano_ingresso, periodo_ingresso, periodos_cursados.

Obteve-se então o *dataset* final com 12 colunas (11 atributos independentes + variável dependente 'perfil') e 216 instâncias, que se referem aos estudantes dos cursos técnicos subsequentes nas situações de concluídos ou evadidos que ingressaram no *campus* de 2017 a 2019.

Na preparação dos dados, todos os valores categóricos foram discretizados em valores numéricos (inteiros). Poucas instâncias precisaram ser ajustadas em relação aos campos nulos. A inclusão de valores foi realizada diretamente nos valores dos atributos dessas instâncias que não foram descartadas. Os dados como "zona" e "renda" foram sintetizados a partir da moda (maior quantidade de ocorrências) obtida do conjunto de dados, sendo esses valores gerados a partir de cada ano letivo de ingresso da qual cada instância fazia parte.

No caso do atributo que representa quantidade de períodos cursados, este foi gerado com uma função de estrutura condicional para obtê-la, pois não havia essa informação no *dataset*.

A variável cidade, oriunda do *dataset* inicial, foi convertida para uma medida de distância geométrica, que resulta na distância em quilômetros que o aluno possivelmente deve percorrer até a cidade de Cajazeiras (PB). O fato de cerca de 2/3 dos estudantes residirem em outras cidades e, por esta razão, na maioria das vezes, precisarem se deslocar diariamente para o Campus, foi um motivo relevante e instigante para a criação deste atributo. Para gerá-lo, foram usados os *shape files* obtidos do portal do Instituto Brasileiro de Geografia e Estatística (IBGE)[1] que contém as geometrias de todas as cidades do Brasil.

O Sistema de Gerenciamento de Banco de Dados PostgreSQL junto à extensão do Postgis foram usados para implementar os *scripts* de criação de um *dataset* que armazena a cidade e a respectiva distância em quilômetros até a cidade de Cajazeiras (PB). A Figura 3 ilustra a *stored procedure* que retorna o cálculo da distância entre as cidades que os estudantes residem até Cajazeiras (PB).

1 Disponível em: https://www.ibge.gov.br/geociencias/downloads-geociencias.html. Acesso em: 23 set. 2022.

Figura 3 ▶
Stored procedure para calcular distância geométrica entre as cidades.
Fonte: arquivo dos autores

```
CREATE OR REPLACE FUNCTION getGeoDistance(VARCHAR)
RETURNS DOUBLE PRECISION
AS $$
    DECLARE
        geo INT;
    BEGIN
        SELECT INTO geo ST_Distance(ST_centroid(a.geom),
                    ST_centroid(b.geom)) * (40075/360) as distancia
        FROM (br_municipios_2019 a JOIN br_municipios_2019 b
              ON (a.nm_mun = 'Cajazeiras' AND b.nm_mun = $1));
        RETURN geo;
    END $$
LANGUAGE plpgsql;

-- Exemplos
SELECT getGeoDistance('João Pessoa'),
    getGeoDistance('Patos'),
    getGeoDistance('Lavras da Mangabeira'),
    getGeoDistance('Sousa'),
    getGeoDistance('Cajazeiras');
```

Data Output | Explain | Messages | Notifications

	getgeodistance double precision	getgeodistance double precision	getgeodistance double precision	getgeodistance double precision	getgeodistance double precision
1	408	136	53	40	0

Foi criado um *dataset* chamado *db_ibge_geo_br* no SGBD mencionado para tratamento e geração das distâncias entre todas as cidades nas quais os estudantes residiam. A tabela *cidades_ifcz* contém três atributos: **id** (chave sequencial); **cidade_origem** (cidade em que reside o estudante, valores extraídos diretamente do relatório da DDE) e **distancia_km** (que armazena a distância em quilômetros a partir do retorno da função *getGeoDistance*).

Ao ser invocada através de uma instrução SQL de consulta, a função *getGeoDistance(varchar)* recebe como parâmetro o nome da cidade em que o estudante reside e consulta na tabela *br_municipios_2019* (oriunda da base de dados do IBGE) por meio de uma equijunção. Se localizada a cidade corretamente, a função calcula a distância geométrica através da função *ST_Distance*, que retorna a distância mínima cartesiana bidimensional entre os dois pontos geométricos da cidade de origem do estudante até Cajazeiras (PB). O valor inteiro encontrado é armazenado no atributo 'distancia_km' da tabela *cidades_ifcz*.

A Figura 3 também mostra alguns testes realizados a partir da execução de chamada à função definida passando algumas cidades como parâmetro, obtendo-se, como exemplos, os valores de distância em relação à cidade de Cajazeiras (PB) para: João Pessoa (408); Patos (PB) (136); Lavras da Mangabeira (CE) (53); Sousa (PB) (40); Cajazeiras (PB) (0). O valor 0 (zero) indica que o estudante reside nesta cidade.

Todos os valores do conjunto de dados foram discretizados para o tipo numérico.

O atributo renda, por exemplo, foi convertido para valores inteiros contínuos a partir do intervalo constante no relatório da DDE, cuja Renda Familiar *Per Capita* (RFP) era definida pelos intervalos de valores definidos na Tabela 1. Essa tabela aponta os valores possíveis de cada um dos atributos selecionados no conjunto de dados para a tarefa de classificação objetivo deste trabalho.

Tabela 1 ▶
Atributos selecionados para tarefa de classificação.
Fonte: dados da pesquisa

#	Atributo	Valores possíveis
0	curso	[0,1] # edificações, eletromecânica
1	distância	[0..408] # mínimo..máximo
2	zona	[0,1] # rural, urbana
3	renda	[0,1,2,3,4,5] # 0<RFP<=0,5; 0,5<RFP<=1; 1<RFP<=1,5; 1,5<RFP<=2,5; 2,5<RFP<=3,5; RFP>3,5
4	cota	[0,1] # não, sim
5	idade	[19..58] # mínimo..máximo
6	sexo	[0,1] # F, M
7	cre	[0..94] # mínimo..máximo
8	ano_ingresso	[2015,2016,2017,2018] # mesmos valores
9	periodo_ingresso	[1,2] # mesmos valores
10	periodos_cursados	[2,3,4,5,6,7] # mesmos valores
11	perfil	[0,1] # não, sim

As duas classes mantiveram a proporção aproximada de 43% (93 instâncias) das classes de estudantes sem perfil de evasão e 57% (123 instâncias) com perfil de evasão (CASTRO; BRAGA, 2011).

Essa alta taxa de evasão e proporção identificadas são bem peculiares aos cursos subsequentes, conforme apresentado na seção introdutória deste trabalho. Essa realidade é bem diferente quando comparada às modalidades de cursos integrados e superiores do *campus*, cuja evasão é bem menor, dentro do intervalo dos anos de 2017 a 2019.

3.3 Modelagem

A plataforma Anaconda[2] foi utilizada para a implementação dos modelos de classificação deste trabalho. Ela dispõe de todas as ferramentas que foram utilizadas nesse projeto; entre elas, o Spyder Python IDE, que disponibiliza recursos desde os básicos aos mais avançados e que integram diversas bibliotecas, como *scikit-learn*, *pandas* e *matplotlib*, para a implementação dos algoritmos e visualização dos resultados. Além das bibliotecas e pacotes disponibilizados, é *open source*, gratuito e possui uma interface simplificada para edição e visualização voltada a conteúdos científicos de dados. Para o estudo, foi escolhida a linguagem de programação Python, que disponibiliza diversos recursos para tarefas de AM.

Após a etapa de pré-processamento dos dados realizada, buscou-se a obtenção do modelo de AM a ser utilizado. Por se tratar de um estudo voltado a um processo de classificação de estudantes quanto ao potencial de evasão, com a abordagem de aprendizado supervisionado e, baseando-se nos resultados dos trabalhos relacionados, alguns algoritmos de árvore de decisão foram escolhidos para serem utilizados e avaliados: *Decision Tree*, *Random Forest* e *Gradient Boosting*. Optou-se por usar *ensembles* também, visando a melhorar o desempenho dos algoritmos de árvores de decisão (FREITAS; GOUVEIA; SOARES, 2020; FRIEDMAN, 2001).

Buscando comparar mais ainda os resultados e avaliar os modelos gerados, foram utilizados outros dois classificadores: KNN e SVM. Esses algoritmos lidam com os valores dos atributos de forma mais independente, diferentemente do que ocorre com as árvores de decisão.

2 Disponível em: https://www.anaconda.com/

4 Resultados e discussão

Nesta seção, são apresentados os resultados obtidos por cada um dos algoritmos experimentados nos quais foram gerados os modelos.

Há duas estratégias principais de particionamento dos dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015): o *hold-out* e a validação cruzada (tradução para *cross-validation*). Segundo Silva, Peres e Boscarioli (2016), a estratégia *hold-out* realiza a criação de dois subconjuntos de dados disjuntos, a partir do *dataset* disponível para uso na indução do modelo, sendo o particionamento realizado uma única vez. Um subconjunto é usado para treinamento (indução) do modelo preditivo, e o segundo para teste após o término do treinamento.

Já no *cross-validation* todos os exemplares farão parte, em algum momento, do conjunto de dados usado no teste do modelo preditivo. Para implementar essa situação, o *dataset* é dividido em K subconjuntos disjuntos, com alocação aleatória de exemplares para cada subconjunto, que ocorre de acordo com o valor definido para k (*k-folds*). O resultado se torna mais confiável, pois geralmente obtém-se o valor médio das *k*-iterações (SILVA; PERES; BOSCARIOLI, 2016).

O valor de *K* foi definido em 10 partições para os algoritmos. Após a execução de várias rodadas de testes com valores para *k*, o valor 10 foi considerado relevante para uma boa estimativa de erro. Essa configuração foi utilizada em todos os algoritmos avaliados.

Não existe uma regra que indique a proporção mais adequada entre os dados de treino teste, todavia observa-se que na maioria dos trabalhos encontrados o tamanho definido para cada parte é numa proporção de 30% para testes de 70% para treino (SILVA; PERES; BOSCARIOLI, 2016). Logo, essa mesma proporção foi utilizada neste trabalho.

Para a avaliação e comparação dos resultados obtidos pelos algoritmos, foram utilizadas algumas métricas, entre elas (MATOS *et al*., 2009): acurácia (ACC), precisão (PRE), recall (REC), f1-score (FS) e suporte (SUP), além da apresentação da AUC. Os valores de métricas obtidos por cada algoritmo são apresentados na Tabela 2, a partir da validação cruzada. A classe positiva (1) corresponde aos estudantes evadidos, e a outra classe (0) representa os estudantes que não evadiram do curso.

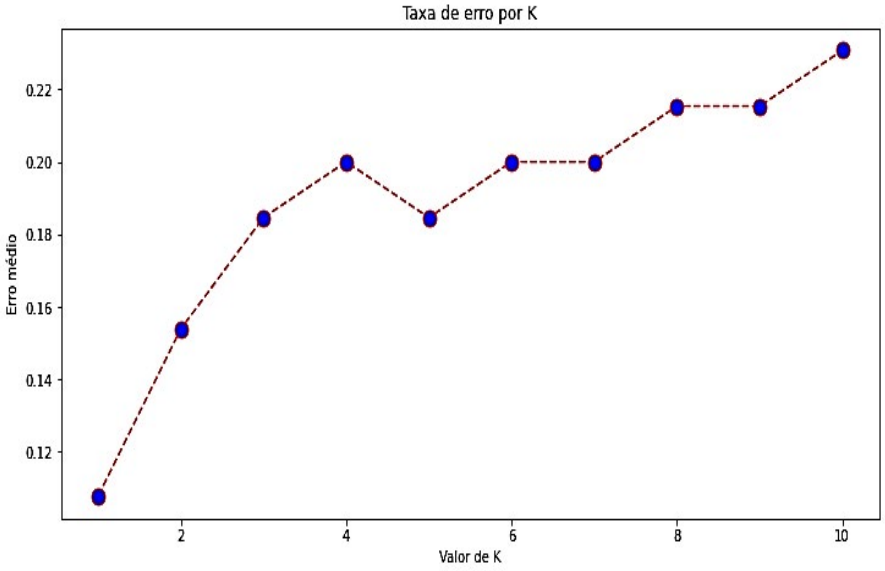
Tabela 1 ▶
Relatório de classificação obtido com os algoritmos.
Fonte: dados da pesquisa

Algoritmo	Classe	PRE	REC	FS	SUP	ACC	AUC
KNN	0 – Não	0.95	0.74	0.77	27	0.78	0.80
	1 – Sim	0.98	0.87	0.85	38		
SVM	0 – Não	0.88	0.78	0.82	27	0.93	0.85
	1 – Sim	0.85	0.92	0.89	38		
Decision Tree	0 – Não	0.89	0.97	0.93	34	0.93	0.92
	1 – Sim	0.96	0.87	0.92	31		
Random Forest	0 – Não	0.87	0.90	0.89	30	0.90	0.89
	1 – Sim	0.91	0.89	0.90	35		
Gradient Boosting	0 – Não	0.90	0.93	0.92	30	0.91	0.92
	1 – Sim	0.94	0.91	0.93	35		

Com esses levantamentos, pode-se considerar que o modelo implementado com KNN obteve um bom resultado tanto para acurácia quanto para AUC.

Na execução de uma das amostragens, a taxa de erro em 10 iterações de *k* mostrou ser crescente conforme apresentado na Figura 4. Observa-se que o erro foi aumentando e se manteve acima de 0.18 a partir da terceira iteração.

Figura 4 ▶
Taxa de erro médio com *k* =10 conforme KNN.
Fonte: dados da pesquisa



O modelo implementado com o SVM obteve melhores resultados em todas as métricas, comparadas aos resultados do KNN. A acurácia obtida com o SVM foi bem superior (93%), bem como a AUC (85%). Observa-se que o KNN foi melhor em relação ao SVM apenas na precisão. Um dos pontos positivos em relação ao KNN é que ele apresenta excelente precisão quando aplicado em tarefas de classificação.

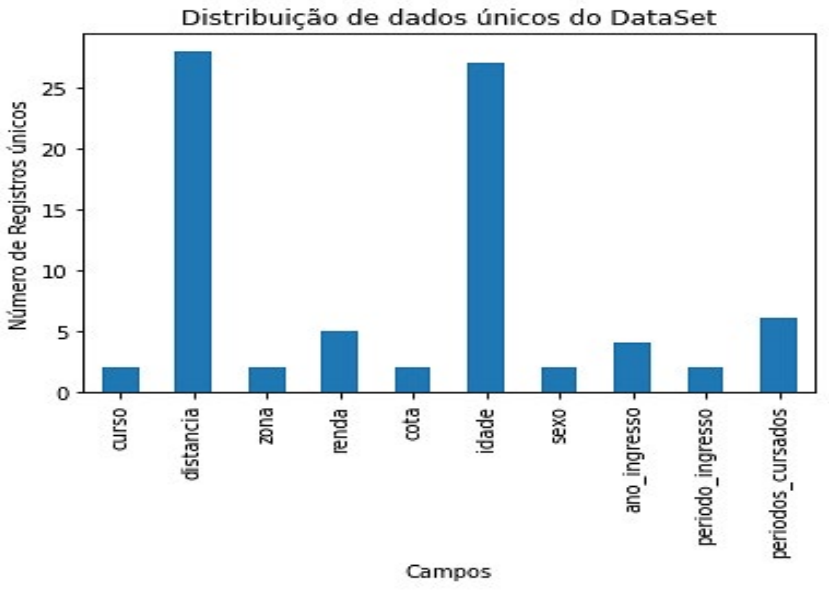
Já os três algoritmos de árvores utilizados apresentaram resultados muito próximos entre eles. Optou-se por alterar alguns dos parâmetros padrão dos algoritmos de árvores, pois observou-se melhor resultado com eles. Assim, o número de árvores geradas foi definido em 100, e a profundidade em 3.

No *Decision Tree*, a classe positiva obteve 96% de precisão, errando apenas uma predição. O *Random Forest* errou um pouco mais em ambas as classes, mas mesmo assim predisse as classes (evadidos e não evadidos) corretamente, obtendo 91% de precisão na classe majoritária e em torno de 90% de *recall* e *f1-score*.

O *Random Forest* obteve, ainda que pouco, um melhor desempenho sobre o *Decision Tree*. O número reduzido de instâncias pode justificar essa diferença de resultados entre os dois algoritmos. Outro fator importante é que o algoritmo calculou a relevância de todos os atributos tratando cada um com alguma percentagem de importância, ainda que baixa para alguns atributos. Percebeu-se, então, que nos testes realizados antes dos resultados aqui apresentados, três atributos foram mais determinantes: 'CRE', 'periodos_cursados' e 'ano_ingresso'. Porém, o atributo 'CRE' estava enviesando os resultados, por isso foi retirado. Com o 'CRE', os algoritmos estavam mostrando resultados altos, por exemplo, com o valor da acurácia maior que 98%, ou seja, com valores determinísticos. A Figura 5 mostra a distribuição dos valores únicos em cada um dos atributos do conjunto de dados, sem o 'CRE'. Os atributos 'distancia' (28) e 'idade' (27) apresentam uma maior

quantidade de variação de valores únicos. Já os demais atributos possuem uma variação menor que 7 valores.

Figura 5 ▶
Distribuição de dados únicos do *dataset*.
Fonte: dados da pesquisa



O *Random Forest* definiu também o atributo 'periodos_cursados' como sendo mais relevante do que o atributo 'ano_ingresso'. Ele errou um pouco mais em ambas as classes, mas mesmo assim predisse as duas classes corretamente com em torno de 90% de *recall* e *f-score*.

Em relação aos algoritmos de árvore de decisão, o *Gradient Boosting* foi o que menos predisse incorretamente ambas as classes e também apresentou excelentes valores para acurácia (91%) e AUC (0.92). O *Gradient Boosting* tratou como relevante praticamente todos os atributos, mas manteve 'periodos_cursados' e 'ano_ingresso' como sendo mais importantes para a classificação. Isso ocorre porque os estudantes que evadem, geralmente estão cursando os 1º e 2º períodos. E os que não evadem tendem a continuar no curso e finalizam no 3º ou (comumente) no 4º período cursado. A Tabela 3 indica esses valores.

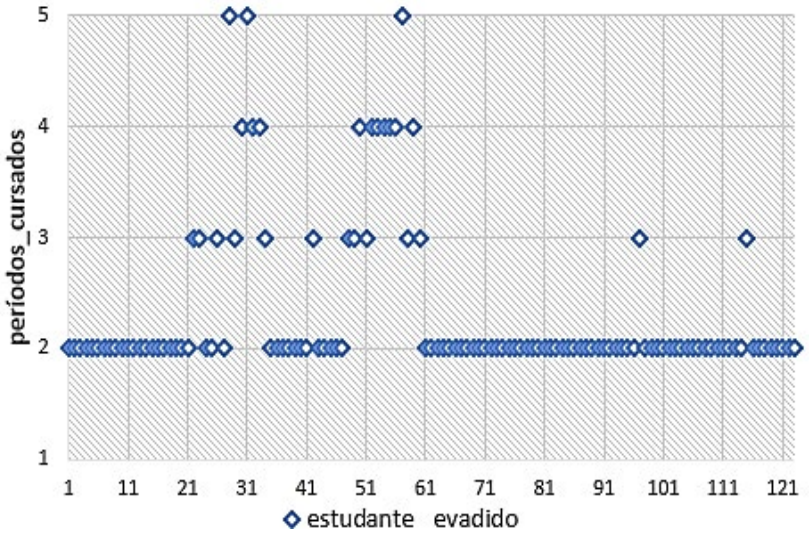
Tabela 3 ▶
Taxa de relevância dos atributos com o *Gradient Boosting*.
Fonte: dados da pesquisa

Atributo	Relevância
periodos_cursados	79.87305 %
ano_ingresso	14.84468 %
idade	2.14422 %
distancia	1.64567 %
cota	0.64736 %
periodo_ingresso	0.44169 %
renda	0.29686 %
zona	0.10947 %
sexo	0.00000 %
curso	0.00000 %

Realizou-se um cálculo para avaliar a relevância em destaque do atributo 'periodos_cursados' e saber a quantidade de períodos cursados pelos estudantes evadidos desde o seu período de ingresso até o período de evasão. A aplicação do método estatístico

'moda' mostra que grande parte dos estudantes evadidos cursaram até 2 semestres letivos, o equivalente a 78,9%, conforme é mostrado na Figura 6. As colunas representam os 123 estudantes evadidos e a respectiva quantidade de períodos cursados, variando de 0 a 5 períodos.

Figura 6 ▶
Quantidade máxima de períodos cursados por estudantes evadidos.
Fonte: dados da pesquisa



No geral, a partir dos valores obtidos pelas métricas apresentadas, percebe-se que em todos os algoritmos os resultados foram considerados muito bons, principalmente aqueles resultados oriundos dos algoritmos de árvore decisão, com mais destaque para o *Gradient Boosting*.

A acurácia foi considerada satisfatória em todos os modelos, à exceção do KNN que ficou com um resultado menor, em torno de 78%. Mesmo assim, ele ainda se destaca como um bom resultado.

Para os estudantes evadidos, numa média geral, os algoritmos conseguiram precisar em cerca de 90% a classe alvo (estudantes evadidos – 1). Os algoritmos também predisseram bem a outra classe de estudantes (não evadidos – 0), porém os valores de *f1-score* mostraram que, no geral, os modelos são um pouco melhores na classe alvo.

5 Considerações Finais

A EDM pode ser considerada uma área de estudo em constante crescimento. Por isso, faz-se necessário que pesquisas sejam desenvolvidas a fim de que se obtenham resultados que favoreçam à área da Educação, ampliando, assim, as tomadas de decisões nas instituições de ensino.

Algumas experiências estão sendo realizadas e investigadas a fim de validar fatores que, à luz dos recursos de EDM, sejam utilizados de modo mais efetivo no tratamento do problema da evasão escolar (RIGO *et al*., 2014).

Neste estudo, foram aplicados alguns algoritmos de AM na identificação de estudantes com risco de evasão escolar usando a tarefa de classificação, com base nos dados obtidos do SUAP e analisados com apoio de um especialista de domínio.

Os dados utilizados na elaboração do modelo se referem a estudantes dos cursos técnicos subsequentes do Campus Cajazeiras do IFPB, entre o período de 2017 a 2019.

Os resultados apresentados pelos algoritmos usados, com taxas de acerto maiores que 80%, indicam que esta é uma abordagem complementar viável para a detecção de estudantes em risco de evasão.

Este trabalho mostrou que a quantidade de períodos cursados é o fator principal para o estudante evadir. Outro ponto a salientar é que a distância geométrica, de onde o estudante reside até o Campus Cajazeiras, não é um fator relevante para a evasão. Isso é importante, pois a maioria dos estudantes são oriundos de outras cidades polarizadas por Cajazeiras.

No *dataset* gerado não havia dados de desempenho acadêmico dos estudantes detalhados por notas semestrais de cada disciplina. Considerando ser este um parâmetro relevante quanto ao índice de evasão dos estudantes, principalmente nos semestres iniciais, um estudo com esses novos dados agregaria insights complementares com maiores detalhes como, por exemplo, a busca por alguns outros motivos da evasão dos estudantes no IFPB, Campus Cajazeiras.

É possível estender o estudo deste trabalho com a criação de novos modelos a serem gerados a partir de outros cursos do Campus Cajazeiras, que apresentam o índice de evasão menor do que os cursos subsequentes, para que, assim, se possa comparar e analisar os resultados obtidos a fim de avaliar a eficiência do modelo proposto em outros cenários. Adicionalmente, o trabalho pode ser ampliado considerando todos os *campi* do IFPB.

De outro modo, outros estudos sobre potencial para evasão poderão ser ampliados e aplicados a partir do cenário educacional gerado em consequência da pandemia.

Financiamento

Esta pesquisa não recebeu financiamento externo.

Conflito de interesses

Os autores declaram não haver conflito de interesses.

Referências

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, p. 3-13, ago. 2011. DOI: http://dx.doi.org/10.5753/rbie.2011.19.02.03.

BARRETO, D. L.; MATOS, M. R.; HORA, H. R. M.; VASCONCELOS, A. P. V. Evasão no ensino superior: investigação das causas via mineração de dados. **Educação Profissional e Tecnológica em Revista**, v. 3, n. 2, p. 3-21, 2019. DOI: https://doi.org/10.36524/profept.v3i2.432.

BÓBÓ, M.; CAMPOS, F.; STROELE, V.; DAVID, J.; BRAGA, R. Identificação do perfil emocional do aluno através de análise de sentimento: combatendo a evasão escolar.

In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE 2019), 30., 2019, Brasília. **Anais** [...]. Brasília: Sociedade Brasileira de Computação, 2019. p. 1431-1440. DOI: http://dx.doi.org/10.5753/cbie.sbie.2019.1431.

BRANCO, U. V. C. Ensino superior público e privado na Paraíba nos últimos 15 anos: reflexões sobre o acesso, a permanência e a conclusão. **Avaliação: Revista da Avaliação da Educação Superior**, Campinas, v. 25, n. 1, p. 52-72, 2020. DOI: http://dx.doi.org/10.1590/s1414-40772020000100004.

BRASIL. Ministério da Educação. **Plataforma Nilo Peçanha:** Guia de referência metodológica – PNP 2020. 2020. Brasília, DF: Evobiz, 2018. Disponível em: http://dadosabertos.mec.gov.br/images/conteudo/pnp/2020/grm_pnp_2020.pdf. Acesso em: 21 jun. 2022.

BRASIL. Ministério da Educação. Secretaria de Educação Superior. Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior. *Associação Brasileira dos Reitores das Universidades Estaduais e Municipais*. Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras. **Diplomação, retenção e evasão nos cursos de graduação em instituições de Ensino Superior públicas**. Brasília, 1996. Disponível em: https://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em: 21 jun. 2022.

CASTRO, C. L.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sociedade Brasileira de Automática: Controle & Automação**, Campinas, v. 22, n. 5, p. 441-466, out. 2011. DOI:https://doi.org/10.1590/S0103-17592011000500002.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. Chicago: SPSS Inc., 2000. Disponível em: http://www.statoo.com/CRISP-DM.pdf. Acesso em: 21 jun. 2022.

CORDEIRO, R. G.; MUSSA, M. S.; HORA, H. R. M. Comportamento de estudantes evadidos de cursos técnicos: um estudo utilizando técnicas de mineração de dados. **Educação Profissional e Tecnológica em Revista**, v. 3, n. 1, p. 87-107, 2019. DOI: https://doi.org/10.36524/profept.v3i1.379.

COSTA, E.; BAKER, R. S. J.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação (JAIE 2012)**, v. 2, p. 1-29, 2012. Disponível em: http://ojs.sector3.com.br/index.php/pie/article/view/2341/0. Acesso em: 21 jun. 2022.

FREITAS, C. N. C.; GOUVEIA, R. M. M.; SOARES, R. G. F. Métodos de machine learning aplicados no cenário da educação a distância brasileira. *In*: INTERNATIONAL SYMPOSIUM ON COMPUTERS IN EDUCATION (SIEE 2020), 22., 2020, On-line. **Proceedings** […]. 2020. Disponível em: http://ceur-ws.org/Vol-2733/paper20.pdf. Acesso em: 21 jun. 2022.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **Annals of Statistics**, v. 29, n. 5, p. 1189-1232, Oct. 2001. DOI: https://dx.doi.org/10.1214/aos/1013203451.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining**: conceitos, técnicas, algoritmos, orientações e aplicações. 2. ed. Rio de Janeiro: Elsevier, 2015.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. **Revista Brasileira de Informática na Educação**, v. 22, n. 1, p. 45-55, abr. 2014. DOI: http://dx.doi.org/10.5753/rbie.2014.22.01.45.

GUERRA, L. C. B.; FERRAZ, R. M. C.; MEDEIROS, J. P. Evasão na educação superior de um Instituto Federal do Nordeste brasileiro**. Revista Eletrônica de Educação (REVEDUC**), v. 13, n. 2, p. 533-553, 2019. DOI: https://dx.doi.org/10.14244/198271992529.

IBM CORPORATION. **IBM SPSS Modeler CRISP-DM Guide**. 2017. Disponível em: https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=spss-modeler-crisp-dm-guide. Acesso em: 21 jun. 2022.

MACHADO, C. J. R.; LIMA, B. R. B.; MACIEL, A. M. A.; RODRIGUES, R. L. An investigation of students behavior in discussion forums using Educational Data Mining. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING & KNOWLEDGE ENGINEERING (SEKE 2018), 30., 2018, Redwood City. **Proceedings […].** Redwood City: KSI Research, 2018. DOI: http://dx.doi.org/10.18293/SEKE2016-143.

MATOS, P.; LOMBARDI, L. O.; CIFERRI, R. R.; PARDO, T. A. S.; CIFERRI, C. D. A.; VIEIRA, M. T. P. **Relatório técnico "métricas de avaliação"**. Projeto "um ambiente para análise de dados da doença anemia falciforme". São Carlos: USP; UFSCar; Unimep, 2009. Disponível em: https://sites.icmc.usp.br/taspardo/techreportufscar2009a-matosetal.pdf. Acesso em: 22 jun. 2022.

MEDEIROS, L. B. G.; PADILHA, T. P. P.; Mineração de dados para detectar evasão escolar utilizando algoritmos de classificação. *In*: CONGRESSO INTERNACIONAL DE EDUCAÇÃO E TECNOLOGIAS; ENCONTRO DE PESQUISADORES EM EDUCAÇÃO A DISTÂNCIA, 4., 2018, São Carlos. **Anais CIET:EnPED:2018 – Educação e Tecnologias: Gestão e política**. São Carlos: UFSCar, 2018. Disponível em: https://cietenped.ufscar.br/submissao/index.php/2018/article/view/623. Acesso em: 10 out. 2020.

PEREIRA, M. C. Evasão escolar: causas e desafios. **Revista Científica Multidisciplinar Núcleo do Conhecimento**, ano 4, ed. 2, v. 1, p. 36-51, fev. 2019. Disponível em: https://www.nucleodoconhecimento.com.br/educacao/evasao-escolar. Acesso em: 21 jun. 2022.

QUEIROGA, E. M. **Geração de modelos de predição para estudantes em risco de evasão em cursos técnicos a distância utilizando técnicas de mineração de dados**. 2017. Dissertação (Mestrado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2017. Disponível em: http://repositorio.ufpel.edu.br:8080/handle/prefix/3843. Acesso em: 21 jun. 2022.

RAMOS, J. L. C.; RODRIGUES, R. L.; SILVA, J. C. S.; OLIVEIRA, P. L. S. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Online. **Anais [...].** Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1092-1101. DOI: https://doi.org/10.5753/cbie.sbie.2020.1092.

RIGO, S.; CAMBRUZZI, W.; BARBOSA, J. L. V.; CAZELLA, S. C. Aplicações de mineração de dados educacionais e *learning analytics* com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, v. 22, n. 1, p. 132-146, abr. 2014. DOI: https://dx.doi.org/10.5753/RBIE.2014.22.01.132.

ROMERO, C.; ROMERO, J. R.; VENTURA, S. A survey on pre-processing educational data. *In*: PEÑA-AYALA, A. (ed). **Educational data mining**: applications and trends. Cham, Switzerland: Springer, 2014. p. 29-44. DOI: https://doi.org/10.1007/978-3-319-02738-8_2.

SANTOS, F. D.; BERCHT, M.; WIVES, L. Classificação de alunos desanimados em um AVEA: uma proposta a partir da mineração de dados educacionais. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE 2015), 4., 2015, Maceió. **Anais** [...]. Maceió: SBC, 2015. p. 1052-1061. DOI: http://dx.doi.org/10.5753/cbie.sbie.2015.1052.

SILVA, J.; DIAS, P.; SILVA, M. C. Fatores de influência no processo de evasão escolar em três cursos técnicos do Instituto Federal de Educação, Ciência e Tecnologia de Brasília. **Revista da UIIPS: Unidade de Investigação do Instituto Politécnico de Santarém**, v. 5, n. 3, p. 6-21, 2017. DOI: https://doi.org/10.25746/ruiips.v5.i3.14522.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados:** com aplicação em R. Rio de Janeiro: Grupo Gen, 2016.