UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Ítalo Vinícius do Rego Lemos

Prevendo a evasão escolar em uma Instituição de Ensino Técnico utilizando Mineração de Dados Educacionais

Recife

2021

Ítalo Vinícius do Rego Lemos

Prevendo a evasão escolar em uma Instituição de Ensino Técnico utilizando Mineração de Dados Educacionais

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciência da Computação

Orientador: André Câmara

Recife

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

L557p Lemos, Ítalo Vinícius do Rego
Prevendo a evasão escolar em uma Instituição de Ensino Técnico utilizando Mineração de Dados Educacionais / Ítalo Vinícius do Rego Lemos. - 2021.
44 f. : il.

Orientador: Andre Camara.
Inclui referências e apêndice(s).

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2021.

1. Mineração de Dados Educaionais. 2. Evasão Escolar. 3. Machine Learning. 4. CRISP-DM. I. Camara, Andre, orient. II. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Ítalo Vinícius do Rego Lemos às 16 horas do dia 16 de julho de 2021, no link meet.google.com/oot-jsvk-jnq, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado Prevendo a evasão escolar em uma Instituição de Ensino Técnico utilizando Mineração de Dados Educacionais, orientado por André Câmara e aprovado pela seguinte banca examinadora:

André Câmara
DC/UFRPE

Rafael Ferreira Leite de Mello
DC/UFRPE

Agradecimentos

Agradeço primeiramente a toda minha família, em destaque para meus pais, Anselmo e Neuma por terem me proporcionado amor e apoio incondicional. E por não medirem esforços em propiciar uma boa educação para mim e minha irmã Nathália. Também agradeço a minha namorada Laís, por ter me incentivado bastante nesta reta final.

Ao meu orientador André Câmara, que me ajudou e me motivou na criação deste trabalho.

Agradeço aos professores do Departamento de Computação da Universidade Federal Rural de Pernambuco que contribuíram para minha formação acadêmica.

A todos amigos e colegas de turma, pelo companheirismo.

Ao Instituto Federal de Educação de Pernambuco, em especial todos os meus colegas de trabalho da Diretoria de Avaliação e Desenvolvimento de Tecnologias, que contribuíram para a realização deste trabalho.

Por último, agradeço a instituição Universidade Federal Rural de Pernambuco, pelas oportunidades conquistadas ao longo do meu período acadêmico e por todo conhecimento obtido.

Resumo

A evasão escolar é um dos principais problemas que ocorrem no âmbito da educação e vem se tornando uma realidade bastante frequente dentro das instituições de ensino públicas ou privadas, resultando em consequências imensuráveis tanto na vida do estudante que deixa de frequentar a escola quanto para a sociedade como um todo. Por ser um fenômeno que preocupa muito os profissionais da educação, se faz necessário revertê-lo, assim estes profissionais necessitam de recursos que sejam eficientes em demonstrar conhecimento dentro e fora do ambiente de ensino e traçar estratégias para lidar com tal cenário. Ser capaz de prever uma possível evasão traz benefícios tando para o estudante quanto para as instituições. A partir disso, uma metodologia que vem se mostrando hábil no combate à evasão escolar e capaz de fornecer conhecimento para a instituição de ensino é a Mineração de Dados Educacionais. Com base nisso, este trabalho teve como objetivo aplicar técnicas de Mineração de Dados e de Aprendizagem de Máquina para prever possíveis casos de evasão antes que o estudante ingresse na instituição de ensino. Através de indicadores sociais e econômicos do estudante e de sua família ele é classificado como um potencial evasor ou não. Este estudo adotou uma base de dados real de uma instituição de ensino pública brasileira, com dados de candidatos que concorreram ao seu processo de ingresso (vestibular) para uma vaga no ensino técnico. Durante a pesquisa foram utilizados 3 modelos de classificação *Decision Tree, Random Forest e XGBoost* tendo o algoritmo *XGBoost* atingindo uma taxa de 74% de acerto na predição de evasores, sendo superior aos demais mas ainda apresentando uma alta número de estudantes classificados como não evadidos mas que se evadiram de fato. Diante desses resultados, concluímos que se faz necessário mais indicadores para detectar, de forma satisfatória, o possível candidato que irá se evadir.

Palavras-chave: Mineração de Dados Educaionais, Evasão Escolar, Machine Learning, CRISP-DM.

Abstract

Dropping out of school is one of the main problems that occur in education and has become a very frequent reality within public or private educational institutions, resulting in immeasurable consequences both in the life of the student who fails to attend school and for society as a whole. Because it is a phenomenon that worries education professionals a lot, if it is necessary to reverse it, so these professionals need resources that are efficient in demonstrating knowledge inside and outside the teaching environment and outlining strategies to deal with such a scenario. Being able to predict a possible dropout benefits both the student and institutions. Based on this, a methodology that has proven to be skillful in combating school dropouts and capable of providing knowledge to the educational institution is Educational Data Mining. Based on this, this work aimed to apply Data Mining and Machine Learning techniques to predict possible dropout cases before the student enters the educational institution. Through social and economic indicators of the student and his family, he is classified as a potential evader or not. This study adopted a real database from a Brazilian public education institution, with data from candidates who competed in its admission process (entrance exam) for a place in technical education. During the research, 3 classification models were used Decision Tree, Random Forest e XGBoost with the XGBoost algorithm achieving a 74% hit rate in predicting evaders, being superior to the others but still presenting a high number of students classified as not dropouts but who actually dropped out. Given these results, we conclude that more indicators are needed to satisfactorily detect the possible candidate who will drop out.

Keywords: Education Data Mining, School Dropping out, Machine Learning, CRISP-DM.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
API	Application Programming Interface
AUC	Area Under the Curve
AVA	Ambientes Virtuais de Aprendizagem
CEFET	Centro Federal de Educação Tecnológica
CVEST	Comissão de Vestibular e Concursos
DADT	Diretoria de Avaliação e Desenvolvimento de Tecnologias
IF	Instituto Federal de Educação, Ciência e Tecnologia
IFPE	Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
RFEPCT	Rede Federal de Educação Profissional, Científica e Tecnológica
ROC	Receiver Operating Characteristics
SGBD	Sistema Gerenciador de Banco de Dados

Sumário

1 Introdução

A educação é primordial na formação de cidadãos críticos capazes de integrar na sociedade de forma intelectual, humana e qualificada (ARRUDA, 2019). Segundo o artigo 1° da lei de diretrizes e bases da educação nacional brasileira ”a educação abrange os processos formativos que se desenvolvem na vida familiar, na convivência humana, no trabalho, nas instituições de ensino e pesquisa, nos movimentos sociais e organizações da sociedade civil e nas manifestações culturais”(BRASIL, 1996). A partir desse princípio, todo e qualquer sistema educativo deve seguir esse propósito com o objetivo de oferecer um caminho de sucesso na formação desses indivíduos.

Com base nisso, para transformar e fornecer condições necessárias para o desenvolvimento educacional e socioeconômico do Brasil foi necessário um grande incentivo na educação e inovação tecnológica (TÁVORA et al., 2015). Uma das grandes conquistas para a oferta da educação profissional e tecnológica no Brasil foi a criação dos Institutos Federais de Educação, Ciência e Tecnologia (IFs) por meio da Lei nº 11.892/2008 (TÁVORA et al., 2015). Só a partir disso, cada estado brasileiro foi beneficiado com pelo menos um destes IFs, sendo criados a partir do potencial instalado nos Centros Federais de Educação Tecnológica (CEFETs), escolas técnicas e agrotécnicas federais e escolas vinculadas às universidades federais (TÁVORA et al., 2015). Ou seja, essas instituições foram importantes para promover uma grande difusão (interiorização) do ensino em todo o território nacional e consequente aumento no número de vagas nas diversas instituições (OLIVEIRA et al., 2019; TÁVORA et al., 2015).

Por sua vez, o ensino técnico pode ser oferecido em articulação com o ensino regular ou por meio de programas de educação continuada (ARRUDA, 2019). Além disso, esses centros de excelência na oferta do ensino técnico tem por objetivo contribuir com o desenvolvimento regional e local por meio da oferta de vagas em cursos qualificantes, técnicos e educação superior, formando profissionais capazes de identificar problemas e criando soluções técnicas e tecnológicas para o desenvolvimento sustentável com inclusão social (ROSINKE et al., 2020; TÁVORA et al., 2015).

No contexto do ensino técnico há uma elevada taxa de evasão escolar ou abandono dos alunos do seu ciclo acadêmico por diversos fatores. Além disso, os prejuízos desse evento nas instituições de ensino é uma realidade e é uma tarefa desafiadora. De acordo com o Ministério da Educação (MEC) a evasão escolar é definida como sendo a ”saída definitiva do aluno de seu curso de origem, sem concluí-lo”(BRASIL, 1997). Já de acordo com Colpani (2018) a evasão também é considerada

quando os alunos se matriculam e não iniciam o curso.

É importante salientar ainda que a problemática da evasão escolar é notória em diversas instituições de ensino nas mais variadas regiões do Brasil e no mundo todo. Sendo também referido como um fenômeno complexo, que pode ocorrer em diversos níveis (integrando a educação básica ao nível superior) e modalidades de ensino (presencial, semi-presencial e EAD), tanto em instituições públicas quanto privadas (NARCISO et al., 2015). Fora isso, as consequências desse fenômeno na maioria das vezes são imensuráveis, trazendo impactos socioeconômicos, sendo apenas observados a longo prazo.

Encarar esse desafio requer uma compreensão dos problemas subjacentes e um planejamento eficaz para praticar intervenções. No Brasil, um dos primeiros esforços governamentais foi a instituição da ”Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras”que teve como objetivo indicar fundamentos da evasão no país e propor soluções para minimizar os números nas instituições de ensino superior públicas (BRASIL, 1997).

Em 2017, a Secretaria de Educação Profissional e Tecnológica que é vinculada ao Ministério da Educação (Setec/MEC) lançou a ferramenta de dados Plataforma Nilo Peçanha (PNP)[1]. A PNP é um ambiente virtual que tem como objetivo coletar, tratar e publicar os dados oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT). Todos os dados divulgados são relativos às informações do corpo docente, discente, técnico-administrativos bem como os gastos de cada unidade da rede.

Como instituição de ensino, o Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE) vem apresentando consideráveis taxas de evasão. De acordo com a PNP, em 2017 a taxa média nacional de evasão escolar para o ensino técnico foi de 22,4%. Delimitando a busca apenas para o contexto das unidades do IFPE, naquele mesmo ano obteve-se uma taxa de evasão de 24,3% para os cursos técnicos.

A partir do monitoramento dos dados da PNP pode-se verificar o real quantitativo de estudantes do IFPE que se enquadram na situação de abandono/evasão. Notou-se que em 2017, o IFPE, possuía 3.763 matrículas, apenas nos cursos técnicos, em situação de abandono, que é aplicada quando o aluno possui mais de 25% de faltas não justificadas e não há mais possibilidade de frequentar as aulas.

Diante da grande presença de dados gerados na área da educação durante os últimos anos, surge um novo campo de estudo denominado Mineração de Dados Educacionais (MDE) (do inglês *Educational Data Mining*) que vem se destacando na

[1] <http://plataformanilopecanha.mec.gov.br>

análise de dados oriundos de ambientes de educação, assim como na evasão escolar. MDE é a aplicação de técnicas de Mineração de Dados (MD) e Aprendizagem de Máquina em conjuntos de dados associados à educação, e assim, seu objetivo é analisar este tipo de dado de maneira a resolver obstáculos em contextos educacionais (ROMERO; VENTURA, 2010).

1.1 Justificativa

A motivação para esta pesquisa surgiu da oportunidade notada a partir de trabalhos prévios realizados no Brasil e no exterior, que abordam como a mineração de dados pode contribuir na solução do abandono escolar. Notou-se que há uma grande gama de trabalhos que abordam a educação apenas no contexto do Ensino Superior como mostra (ALBAN; MAURICIO, 2019), destacando mais de mil trabalhos inseridos nessa condição.

Além disso, o perfil de um estudante evasor é sempre analisado levando em consideração características do seu pós-ingresso na instituição. Assim, se faz necessário criar modelo que acompanhe o estudante a partir da sua tentativa de ingresso na instituição além de abranger o ensino técnico. Desta forma, será possível facilitar a identificação de futuros casos de evasão pela instituição. O modelo utilizará dados socioeconômicos de vestibulares anteriores e informações do sistema acadêmico.

Também de forma a contribuir com as políticas públicas institucionais direcionadas para a permanência e o êxito escolar em Institutos Federais (ALVAREZ; MATOS, 2020). Tendo conhecimento de todas essas informações, identificou-se a necessidade de conhecer o comportamento dos evadidos antes do seu ingresso na instituição e um processo que possa predizer as reais chances de um estudante evadir-se do curso no decorrer da sua trajetória acadêmica.

Todo o resultado desta pesquisa pode servir de base para novos projetos que venham a ser desenvolvidos sobre o mesmo tema, podendo utilizar o mesmo conjunto de dados, técnicas empregadas e atributos escolhidos.

1.2 Objetivos

Essa seção contém o objetivo geral e os objetivos específicos que serão utilizados para guiar esse trabalho.

1.2.1 Objetivo Geral

O presente trabalho tem como objetivo analisar os dados relativos dos candidatos com risco de evasão utilizando técnicas de Mineração de Dados Educacionais além de investigar possíveis indicadores que levem o candidato a evadir-se após o ingresso numa instituição de ensino.

1.2.2 Objetivos Específicos

- Selecionar as características mais relevantes para compor o banco de dados das amostras;
- Preparar os conjuntos de dados para treinamento e teste do modelo;
- Comparar modelos de classificação aplicados na predição da evasão;
- Analisar os resultados com os modelos utilizados;

1.3 Estrutura do Trabalho

Este trabalho divide-se da seguinte forma. Além do primeiro capítulo onde foi apresentado a introdução, justificativa, objetivos do tema abordado, o trabalho contém outros 5 capítulos. O segundo capítulo explana sobre a fundamentação teórica deste trabalho. No capítulo três, os trabalhos relacionados são apresentados. O capítulo quatro mostra como foi desenvolvida a proposta desse trabalho. O capítulo cinco apresenta os resultados obtidos dos testes realizados, a fim de avaliar a abordagem desenvolvida. No sexto capítulo as conclusões sobre o trabalho são apresentadas, assim como a proposta de trabalhos futuros.

2 Fundamendação Teórica

Este capítulo aborda os conceitos necessários para uma melhor compreensão dos tópicos abordados neste trabalho.

2.1 Evasão escolar

Uma das maiores preocupações de quem está inserido na área educacional é a questão do abandono escolar e evasão escolar. Ambos os termos possuem suas definições correlacionadas, mas se diferenciam em alguns aspectos. Segundo Filho e Araújo (2017) a diversidade de conceituação atrapalha a quantificação precisa dos casos, o que pode dificultar a falta de clareza e objetividade na superação deste problema. Para o INEP, abandono é caracterizado quando o estudante se desliga da escola, de uma disciplina ou curso, mas retorna no ano/semestre posterior. Já a evasão abrange estudantes que deixam de frequentar o ambiente escolar e no ano seguinte não efetuam sua matrícula regularmente, abandonando em definitivo a continuidade dos estudos.

Tanto abandono quanto evasão escolar representam um processo complexo e cumulativo de fatores que levam o estudante a efetuar a sua saída do ambiente educacional. Para Lüscher e Dore (2011), existem um conjunto de fatores que influenciam a evasão, tanto ao estudante e à sua família quanto à escola e à comunidade em que se vive. A autora define os conjuntos através de duas perspectivas, detalhadas a seguir:

- **Perspectiva Individual:** Nela encontram-se os valores, os comportamentos, as atitudes que promovem um maior ou menor engajamento do estudante na vida escolar. Além disso, o background familiar (nível escolar dos pais, renda familiar e estrutura da família) é reconhecido como o fator mais importante para o sucesso ou insucesso do estudante em seu percurso escolar.
- **Perspectiva Institucional:** Esta perspectiva compõe fatores relacionados à evasão ou à permanência do estudante na escola, como composição do corpo docente, os recursos escolares, estrutura física da escola além das práticas pedagógicas adotadas.

Dados relevantes sobre o tema no Brasil podem ser consultados através da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), que é realizada trimestralmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Seu

objetivo principal é produzir indicadores e outras informações necessárias para o estudo socioeconômico do país.

Em 2020 teve pela primeira vez a divulgação de dados sobre abandono escolar. De acordo com o levantamento do PNAD Contínua 2020, ano base 2019, existem o equivalente a 50 milhões de pessoas de idade entre 14 e 29 anos, que correspondem a todo o quantitativo de jovens brasileiros, onde 20,2% (10,1 milhões de jovens) não completaram o ensino médio. Tendo eles abandonado a escola antes do término desta fase ou por nunca tê-la frequentado.

O relatório do PNAD também apontou as causas principais que levaram estes jovens a abandonar ou nunca comparecer à escola. Dentre as principais causas estão a necessidade de trabalhar apontado como o principal motivo, tanto entre homens quanto mulheres, seguidos da falta de interesse em estudar e da gravidez. Além disso, 11% das mulheres apontaram realizar trabalhos domésticos como fator chave para afastar-se da escola.

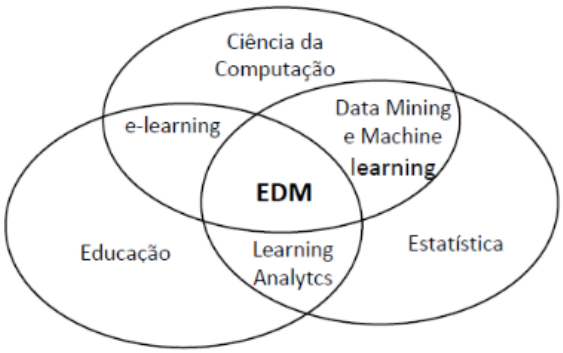
2.2 Mineração de Dados Educacionais

Mineração de Dados Educacionais (MDE) é um campo recente e ativo dentro da ciência da computação. Tendo como principal precursor as estruturas técnicas e ferramentas de Mineração de Dados (MD) em um contexto educacional. De acordo com Romero et al. (2010) MDE é reconhecida como uma disciplina emergente e seu objetivo é analisar, explorar e desenvolver métodos nesses ambientes de modo a encontrar soluções apropriadas para o campo da educação. Entre os resultados esperados estão o aprimoramento do aprendizado dos estudantes, detectar comportamentos e gerar recomendações, avaliar a performance dos professores, organizar os recursos da instituição educadora entre outros.

A MDE ao longo dos últimos anos vem sendo empregada em algumas tarefas como por exemplo, na análise de performance acadêmica dos estudantes na qual identifica se ele irá ser aprovado ou reprovado em determinada disciplina e também na identificação de estudantes propensos à evasão (ASIF et al., 2017; AULCK et al., 2016; MANHÃES et al., 2012; PEREZ; CASTELLANOS; CORREAL, 2018).

Segundo Romero e Ventura (2013) MDE é uma área interdisciplinar que pode ser definida como a combinação de três principais áreas: Ciência da Computação, Educação e Estatística. E através da intersecção destas áreas outras sub-áreas são formadas, conforme a Figura 1.

Figura 1 – Principais áreas relacionadas a MDE

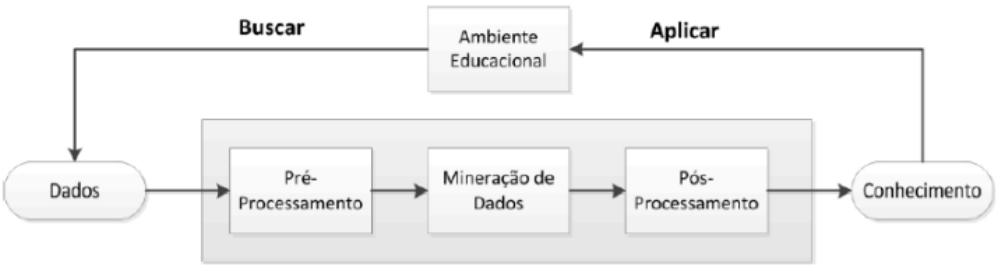


Fonte: (RODRIGUES et al., 2014)

Com a ampla difusão do uso de sistemas informatizados nas escolas e universidades, cresce a cada dia o volume de dados gerados e armazenados em bases de dados produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como Ambientes Virtuais de Aprendizagem (AVA), Sistemas Tutores Inteligentes, entre outros (COSTA et al., 2013; RIGO et al., 2014). Este grande volume de dados tem fomentado o interesse na sua utilização, junto com técnicas de Mineração de Dados, relacionadas com o processo de aprendizagem (RIGO et al., 2014).

O processo MDE converte os dados brutos vindos de sistemas educacionais em conhecimento que pode ser utilizado por estatísticos, professores, desenvolvedores de software etc. (RODRIGUES et al., 2014). As etapas são baseadas nos mesmos passos da mineração de dados, de acordo com a Figura 2.

Figura 2 – Etapas da Mineração de Dados Educacionais



Fonte: (RODRIGUES et al., 2014)

Técnicas de Aprendizado de Máquina (AM) também fazem parte do estudo da MDE. Uma das áreas mais atuantes dentro da MDE é a de predição (BAKER et al., 2010). O objetivo de realizar análises preditivas é inferir um atributo de destino a partir da combinação de outros aspectos dos dados. Os tipos de métodos de uma predição

podem ser classificação, quando se prever uma variável categórica, ou regressão, quando se prever um variável de valor contínuo.

2.3 Aprendizado de Máquina

Aprendizado de Máquina (AM) é um campo de estudo derivado da Inteligência Artificial e da Ciência da Computação que tem em sua essência, habilitar um sistema computacional a aprender e pensar automaticamente, agindo como o pensamento humano. Isso ocorre através do desenvolvimento de algoritmos que acessam dados e informações e aprendem por si próprios.

O processo de conhecimento do AM começa com a observação dos dados a fim de encontrar padrões nestes dados e tomar decisões baseadas nas informações que foram providas.

Em geral, os algoritmos de AP são classificados em três categorias de acordo com o seu paradigma.

- **Aprendizagem Supervisionada:** Essa categoria é definida pelo uso de dados rotulados para treinar modelos e assim predizer resultados futuros com uma maior precisão.
- **Aprendizagem Não Supervisionada:** Aprendizagem Não Supervisionada: Utiliza dados não rotulados, realizando uma análise para descobrir padrões ocultos sem a necessidade de intervenção humana.
- **Aprendizagem Semi-Supervisionada:** Faz uso tanto de dados rotulados como não rotulados durante o treinamento de modelos.

2.3.1 Aprendizagem Supervisionada

Esse tipo de aprendizado é comumente associado aos modelos preditivos, e suas atividades mais comuns são a classificação e a regressão. O modelo de predição funciona com a aplicação de funções para estimar valores desconhecidos ou futuros em função das características das variáveis independentes relacionadas. Ou seja, o modelo preditivo almeja indicar o resultado de uma variável de interesse a partir de valores já conhecidos de outras variáveis.

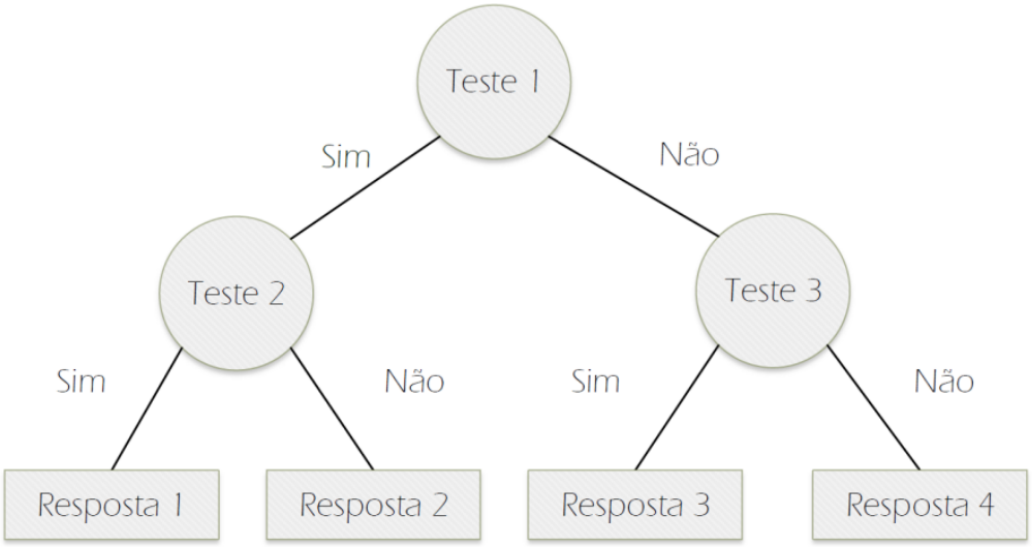
2.4 Árvore de Decisão

Árvore de decisão é uma técnica de AM que possui uma estrutura de árvore semelhante a um fluxograma que mapeia possíveis resultados a partir da avaliação de

atributos e retorna uma predição baseada nos valores desses atributos (PRIYAM et al., 2013). Ela é uma estrutura de dados composta por nós e ramos. Cada nó contém um teste referente a um atributo do dado de entrada e dependendo do resultado ele se ligará a outro nó através de um ramo descendente correspondente ao resultado deste teste. O nó folha, ou o nó terminal, é aquele que não possui mais ramos, ele indica o valor classificado após todos os testes realizados.

Assim, o algoritmo tem seu início a partir do nó raiz, em seguida percorre toda árvores realizando testes sobre cada atributo até chegar ao nó folha, que é onde será definido o resultado. A figura 3 representa a estrutura de uma árvore de decisão binária.

Figura 3 – Estrutura de uma árvore de decisão.



Fonte: (ROLIM V.B, 2014)

A árvore de decisão utilizada neste trabalho é do tipo binária, entre elas podemos citar dois tipos: *Random Forest* e *Gradient Boosting*.

2.4.1 Random Forest

O modelo utilizado em *Random Forest* consiste em uma grande quantidade de árvores de decisão que operam como um conjunto. O resultado da predição, em casos de classificação, é baseada na maioria dos votos dos valores previstos em cada árvore de decisão (RESENDE; DRUMMOND, 2018).

2.5 Rede Federal de Educação Profissional, Científica e Tecnológica

Instituída a partir da Lei 11.892, de 29 de Dezembro de 2008, a Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) se tornou uma organização política, vinculada ao Ministério da Educação, das instituições federais de educação profissional e tecnológica de acordo com o artigo 1 da (BRASIL, Lei 11.892, de 29 de Dezembro de 2008).

Ainda de acordo com o artigo 1 da (BRASIL, Lei 11.892, de 29 de Dezembro de 2008) a RFEPCT é composta pelas seguintes instituições:

I Institutos Federais de Educação, Ciência e Tecnologia - Institutos Federais;

II Universidade Tecnológica Federal do Paraná - UTFPR;

III Centros Federais de Educação Tecnológica Celso Suckow da Fonseca - CEFET-RJ e de Minas Gerais - CEFET-MG;

IV Escolas Técnicas Vinculadas às Universidades Federais; e

V Colégio Pedro II.

Em seu artigo 2 (BRASIL, Lei 11.892, de 29 de Dezembro de 2008) define os Institutos Federais como instituições de educação superior, básica e profissional, pluricurriculares e multicampi, especializados na oferta de educação profissional e tecnológica nas diferentes modalidades de ensino.

2.6 Instituto Federal de Educação de Pernambuco

O IFPE é uma instituição vinculada à Rede de Educação Profissional e Tecnológica, criada em 2008 através da Lei nº 11.892/08, o Instituto oferece uma proposta inédita de ensino verticalizado, articulando, num só lugar, 54 cursos que atendem estudantes em diferentes níveis e modalidades de formação: ensino médio, técnico, superior nas modalidades Tecnológico, Licenciatura e Bacharelado, além de especialização e mestrado.

O instituto possui 16 campi distribuídos entre as mesorregiões do estado, complementado por uma unidade de Educação a Distância, que conta com 11 polos. Atualmente seu processo seletivo abrange duas formas de ingresso. Podendo ser através do SISU (Sistema de Seleção Unificada) para alunos que realizaram o Exame Nacional do Ensino Médio (ENEM) ou através do vestibular tradicional que ocorre duas vezes ao ano, uma em cada semestre.

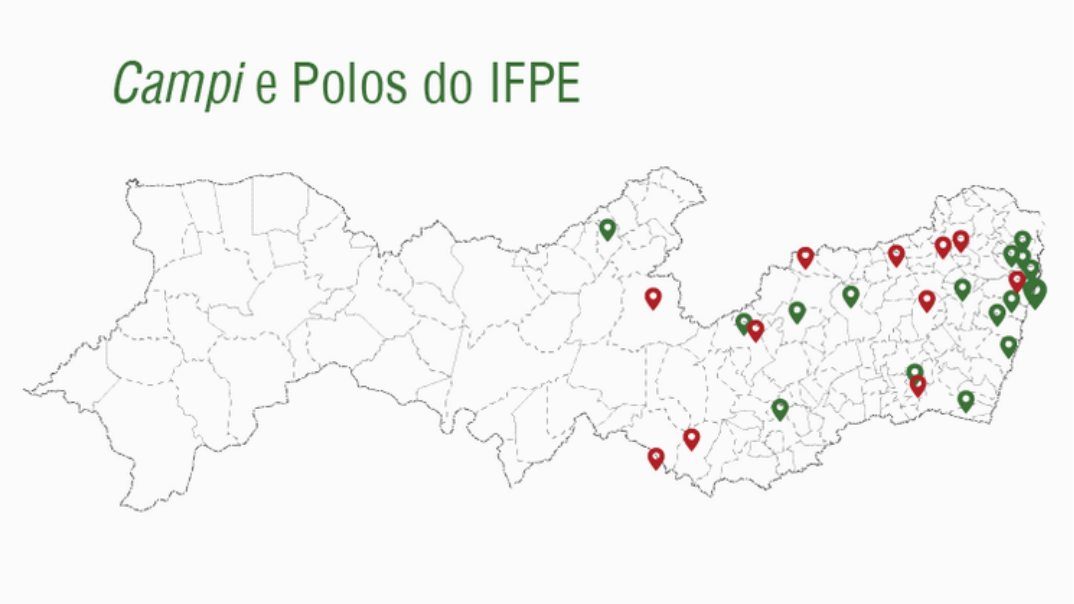


Figura 4 – Campi e Polos do Instituto Federal de Pernambuco

3 Trabalhos Relacionados

Nesta seção serão apresentados trabalhos relacionados ao tema de Mineração de Dados Educacionais em situação de evasão escolar. Eles mostram como o assunto vem sendo tratado e quais técnicas estão obtendo bons resultados e sua importância.

O trabalho de Silva et al. (2019) apresentou um estudo de caso com o uso de dados estatísticos de mais de 40 mil estudantes sobre a previsão da evasão do município de Juiz de Fora/MG. Utilizando a abordagem de descoberta e conhecimento de bases de dados (KDD - *Knowledge Discovery in Databases*) para realizar a mineração de dados, e aplicando o algoritmo floresta randômica ponderada (*Weighted Random Forest*) para classificação. Para tal predição, foi fundamental verificar quais atributos são mais relevantes para classificar o estudante como candidato a evasão. O estudo apresentado concluiu que a abordagem adotada obteve 76% de precisão na possibilidade de prever um aluno com potenciais sinais de se evadir.

Já no estudo de Aulck et al. (2016), foi analisado as características chaves que indicam evasão. Seu foco principal é predizer a evasão do aluno através de informações demográficas bem como registros escolares. Os autores utilizaram a base de dados da universidade de Washington (UW), com dados de estudantes da graduação entre os anos de 1998 a 2006, totalizando cerca de 69 mil registros. Seu trabalho utilizou três algoritmos de classificação (Regressão logística regulariza, KNN e *Random Forest*), obtendo performance diferente para cada algoritmo.

Podemos também identificar no trabalho de Maria, Damiani e Pereira (2016) uma forma de aplicar redes bayesianas na intenção de predizer percentualmente as chances de evasão de estudantes. Este estudo foi realizado com base nas características dos alunos, coletadas do sistema utilizado pelo SENAI/SC. A sua importância se deve ao fato que a partir da validação dos resultados foi verificado uma taxa de 85,6% de acerto, o que demonstrou um bom desempenho da rede bayesiana modelada para o sistema desenvolvido, auxiliando assim, os gestores educacionais com os percentuais de chance de evasão dos alunos.

O trabalho de Barbosa, Santos e Pordeus (2017) discute uma estratégia de predição de evasão através de classificação com o paradigma de opção e rejeição. As instâncias do estudo são classificadas em 3 classes, estudantes propensos à evasão, não propensos à evasão e estudantes que não são classificados, são considerados “rejeitados”. O paradigma evita classificar casos em que não há características suficientes, assim essa instância pode ser reutilizada por outro classificador diferente ou classificada manualmente. Para esse estudos foram utilizados dados de 892 estudantes do curso

de Ciência da Computação da Universidade Federal do Ceará (UFC) matriculados entre os anos de 2005 a 2016. A intenção dos autores é mostrar que os estudantes não-classificados tem chance de terem mais sucesso quando assistidos por educadores para consequentemente conseguirem terminar seus estudos.

Em outro trabalho, Beltran et al. (2019) adotaram o uso de dados acadêmicos e socioeconômicos de alunos de uma universidade de ensino superior do Peru. A partir do pré processamento desses dados escolheu-se um modelo de aprendizado de máquina responsável por informar se um aluno possui ou não risco de evasão. Essa escolha utilizou os seguintes algoritmos: *Naive bayes*, MLP, *AdaBoost*, J48, IBk e *Bagging*. Por fim, o modelo adotado foi agregado a uma plataforma desenvolvida com a intenção de auxiliar alunos alunos em risco de evasão.

Através da metodologia de Tasnim, Paul e Sattar (2019) buscou-se características para calcular o valor mínimo para ser considerado um estudante com risco de se evadir. No cálculo do valor limite os atributos do conjunto de dados são categorizados em duas propriedades: fatores de incremento ou decremento. Ou seja cada atributo terá uma habilidade de aumentar ou diminuir as chances do aluno ser considerado propenso a se evadir. A base utilizada, nomeada de ”Análise de desempenho do estudante”, foi obtida através do repositório UCI *Machine Learning Repository* e contou com cerca de 1.044 amostras.

Existem diversas abordagens para prever o risco de evasão, diferentes modelos utilizados e em diferentes níveis de ensino (Tabela 1). Contudo, durante a análise da literatura atual, todos os trabalhos encontrados analisavam o estudante após o seu ingresso na unidade de ensino, seja ele após cursar o primeiro ano ou após ele cursar todo o curso. Assim utilizam apenas de dados acadêmicos, como por exemplo, notas ou quantidades de reprovações desconsiderando aspectos sociais. Portanto, diferentemente dos trabalhos existentes, nosso estudo aborda o estudante em um contexto pregresso a sua admissão na instituição de ensino, utilizando fatores sociais e econômicos individuais e de sua família, como por exemplo: tipo de moradia, grau de educação dos pais, renda familiar, acesso a internet, entre outros.

Tabela 1 – Tabela de resumo dos trabalhos relacionados.

Trabalho	Modelos de Classificação	Nível de Ensino
(SILVA et al., 2019)	*Weighted Random Forest*	Superior
(AULCK et al., 2016)	KNN, *Random Forest*	Superior
(MARIA; DAMIANI; PEREIRA, 2016)	*Naive Bayes*	Técnico
(BARBOSA; SANTOS; PORDEUS, 2017)	KNN, SVM, *Naive Bayes*, MLP	Superior
(BELTRAN et al., 2019)	*AdaBoost, Bagging*, IBk, J48, *Naive Bayes*, MLP	Superior
(TASNIM; PAUL; SATTAR, 2019)	*Logistic regression*, *Naive Bayes*, SVM	Superior
Trabalho proposto	*DecisionTree, Random Forest, XGBoost*	Técnico

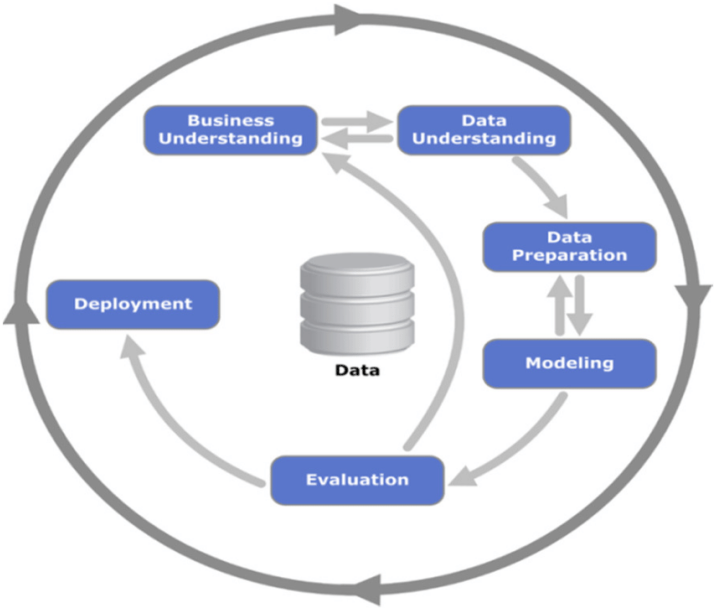
Fonte: O Autor

4 Desenvolvimento

Este capítulo aborda o processo de desenvolvimento desta pesquisa, fornecendo as etapas necessárias para a compreensão de como correu seu desdobramento. A seguir apresentamos a metodologia adotada para esta fase.

Para a análise da evasão foi escolhida a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) (CHAPMAN et al., 2000). Trata-se de um processo bastante utilizado pela Mineração de Dados. Sua vantagem é ser uma técnica que pode ser aplicada a qualquer tipo de negócio e não há restrições de ferramentas ou de tecnologias específicas para ser executada.

Figura 5 – Fases da metodologia CRISP-DM.



Fonte: (CHAPMAN et al., 2000).

O ciclo de vida da metodologia CRISP-DM, como observado na Figura 5, é dividido em seis fases, :

1) Compreesão do Negócio (*Business Understanding*)
2) Entendimento dos Dados (*Data Understanding*)
3) Preparação dos Dados (*Data Preparation*)
4) Modelagem (*Modeling*)
5) Avaliação (*Evaluation*)

6) Implantação (*Deployment*)

Os detalhes sobre cada fase serão esclarecidos nas subseções a seguir.

4.1 Compreensão do Negócio

A etapa inicial visa determinar os objetivos da pesquisa, bem como realizar uma análise da literatura especializada para obter informações sobre a evasão de estudantes e os trabalhos já desenvolvidos anteriormente. Além de converter para um problema de mineração de dados e o caminho trilhado para atingir o objetivo geral. Toda parte de contextualização, objetivo do problema foram tratados no capítulo 1 e 2 deste trabalho. As demais etapas são descritas neste capítulo.

4.2 Entendimento dos dados

A segunda etapa lida com os dados brutos que irão ser utilizados para apoiar a solução do problema, abrangendo a coleta, exploração e qualidade dos dados. Esta etapa tem como objetivo compreender a fundo o conjunto de dados, suas características, tipos de dados, total de registros etc.

4.2.1 Coleta de Dados

Esta seção descreve como os conjuntos de dados para a execução deste trabalho foram obtidos. A coleta pretendeu adquirir os dados referentes às inscrições ocorridas no processo seletivo tradicional do IFPE para aplicação das técnicas de mineração de dados e *machine learning*, de modo a contribuir na predição do abandono escolar.

A base de dados escolhida para ser analisada foi a de processos de ingressos (vestibulares) já ocorridos em todos os campi físicos do IFPE. A razão desta escolha foi devido a possibilidade de obter informações apresentadas pelos candidatos, futuros estudantes, antes de ingressar na instituição.

O processo de coleta teve início através do contato com a Diretoria de Avaliação e Desenvolvimento de Tecnologias (DADT), onde foi apresentada a proposta deste trabalho assim como as informações necessárias para continuar com a pesquisa. Em seguida, a DADT apresentou a proposta para a comissão de vestibular e concursos (CVEST) do IFPE, que é o setor responsável por organizar e administrar o vestibular tradicional que ocorre anualmente na instituição, a fim de obter os dados requeridos. Obtida a autorização para utilização dos dados, foi repassado à DADT um usuário, apenas com permissão de leitura, para ter acesso ao banco de dados institucional. O

Sistema Gerenciador de Banco de Dados (SGBD) utilizado pela CVEST é o MySQL e seu acesso ocorreu através da aplicação web phpMyAdmin.

Para cada seleção de vestibular é gerado um novo banco de dados correspondente ao ano e ao semestre de sua realização. Em seguida, foram extraídas as bases correspondentes aos vestibulares dos anos/semestre 2016.1, 2016.2, 2017.1, 2017.2, 2018.1, 2018.2, 2019.1 e 2019.2. Totalizando 8 conjuntos de dados. Cada conjunto de dados é composto por duas tabelas nomeadas de ***inscricao*** e a de ***resposta_questionario***. Assim, para cada ano/semestre tem-se duas tabelas, totalizando 16 tabelas. Os dados foram obtidos em arquivo de formato ".csv", na qual se iniciou a exploração dos dados. Um resumo da quantidade de dados contidos na base é apresentado na Tabela 2.

Tabela 2 – Dados dos vestibulares do IFPE

Semestre	Quantidade de Inscrições
2016.1	53.857
2016.2	849
2017.1	57.605
2017.2	24.336
2018.1	44.399
2018.2	19.759
2019.1	45.369
2019.2	24.644
Total	1.118,969

Fonte: O Autor

4.2.2 Exploração dos Dados

Esta seção aborda a etapa de exploração dos dados, examina propriedades como formato dos dados, número de registros e campos em cada tabela e seus relacionamentos. A partir desta seção, para leitura, visualização e manipulação dos dados, foi utilizada a linguagem de programação Python, com o conjunto de bibliotecas Pandas[1] e Scikit-Learn[2].

Antes de tudo, é importante conhecer como é o procedimento utilizado na inscrição no processo seletivo de vestibular do IFPE. O processo consiste no candidato realizar uma inscrição em um endereço eletrônico, onde é preenchido um cadastro com informações obrigatórias (nome civil completo, endereço, documento de identificação, curso/turno/entrada, se é cotista ou não), após isso é solicitado o preenchimento de um questionário socioeconômico. O questionário socioeconômico tem como objetivo

[1] <https://pandas.pydata.org/>
[2] <https://scikit-learn.org/>

conhecer o contexto econômico, social e o processo de formação educacional do candidato e seus pais, não influindo na classificação do candidato.

Após o preenchimento do questionário é gerada uma taxa de inscrição para posterior pagamento. Realizado o pagamento a inscrição é efetivada tornando o candidato apto a realizar o exame de seleção. Candidatos isentos da taxa de inscrição já possuem inscrição efetivada.

Em relação às tabelas que compõem cada base de dados temos as seguintes definições. A tabela ***inscricao*** contém atributos referente a cada inscrição efetuada pelos candidatos, contendo dados pessoais (não presentes), indicação do nível escolar a qual desejam concorrer, o curso desejado, o campus escolhido, se é cotista ou não e as notas obtidas após a realização das provas do exame. A tabela contém todas as inscrições realizadas, incluindo aquelas com status de não efetivadas por falta de pagamento. A tabela contém uma coluna que indica qual a situação do candidato após a realização do exame de seleção, podendo ser os seguintes status: APROVADO, REMANEJÁVEL, DESCLASSIFICADO e FALTOU.

Os principais formatos de dados encontrados foram textos, e numéricos apenas para as colunas que armazenam o resultado do exame de seleção.

A segunda tabela apresenta as respostas sobre o questionário socioeconômico aplicado no ato da inscrição. Todo candidato ao se inscrever deve responder este questionário, criando um relacionamento de 1:1 com a tabela inscrição. A tabela ***resposta_questionario*** possui um total de 40 colunas, todas do tipo numérico. A partir das respostas do questionário será montado um perfil de candidatos com perfil de evasão.

4.2.3 Qualidade dos Dados

Esta seção examina a qualidade dos dados, checando se os dados estão completos e se possuem dados faltantes.

Durante a pré-visualização dos dados, foi possível identificar que o conjunto possui uma taxa nula de dados ausentes, isso se deve ao fato de que para realizar a inscrição grande parte dos dados requisitados são de preenchimento obrigatório.

4.3 Preparação dos Dados

Essa etapa inclui a seleção de atributos, limpeza de dados, formatação e transformação para aplicação dos algoritmos de ML na etapa posterior.

4.3.1 Seleção dos dados

Essa seção teve como objetivo a construção da base de dados que foi utilizada na geração dos modelos para aplicação dos algoritmos de ML. Foram selecionados quais atributos são relevantes para o abandono escolar. Ela foi dividida em duas fases para atingir seu resultado. A primeira fase teve o propósito de excluir atributos, filtrar registros necessários para este estudo e a segunda fase objetivou criar.

Primeiramente, cada conjunto de dados das inscrições recebeu uma série de filtros para que atendesse a proposta desta pesquisa. O primeiro filtro aplicado teve a intenção de selecionar apenas inscrições com o status de efetivada, removendo aquelas não aptas à realização das provas. Em seguida, aplicamos o filtro para selecionar apenas inscrições que concorrem a cursos técnicos na modalidade subsequente. Por último, filtramos pela situação do candidato após a realização da prova, sendo apenas consideradas as inscrições com situação de APROVADO e REMANEJÁVEL.

Na tabela ***inscricao***, foram utilizados atributos com características individuais do candidato como sexo, cor, estado civil além da indicação no uso de cota ou não, notas nas provas de português, matemática e conhecimentos gerais e a média final.

A tabela a seguir mostra os atributos selecionados do conjunto ***inscricao*** e seus possíveis valores.

Tabela 3 – Relação dos atributos selecionados da tabela ***inscricao***

Atributo	Possíveis Valores
Cor	BRA - Branca NEG - Preta PAR - Parda AMA - Amarela IND - Indígena NÃO - Não declarada
Sexo	M - Masculino F - Feminino
EstadoCivil	S - Solteiro(a) C - Casado(a) D - Divorciado(a) V - Viúvo(a) SP - Separado(a) O - Outros
Cotista	S - Sim N - Não
Portugues	Valor numérico entre 0 e 100
Matematica	Valor numérico entre 0 e 100
Consgerais	Valor numérico entre 0 e 100

Fonte: O Autor.

Na tabela ***resposta_questionario***, foram selecionados atributos com características socioeconômicas do candidato e de sua família, tais como renda familiar, se o mesmo trabalha, se possui acesso a internet, bem como informações sobre a formação escolar dos seus pais. Todos os atributos selecionados estão presentes na tabela 4 assim como seus possíveis valores de resposta. Os demais atributos foram desconsiderados por não trazerem relevância ao estudo.

Tabela 4 – Perguntas selecionadas do Questionário Socioeconômico

Atributo	Pergunta	Possíveis Valores
q8	Quantos filhos você tem?	Nenhum Um Dois Três Quatro ou mais
q10	Qual sua localização ou região de moradia?	Zona Urbana Zona Rural
q15	Você exerce algum trabalho remunerado?	Não. Sim (Vínculo formal). Sim (Vínculo informal).
q18	A sua família participa de algum programa de transferência de renda?	Bolsa família. BPC (Benefício de Prestação Continuada). Jovem aprendiz. PROJOVEM. PETI. Chapéu de palha. Seguro safra. Outros. Nenhum
q19	Qual a renda mensal do seu grupo familiar?	Até meio salário mínimo. Até 1 salário mínimo. Até 2 salários mínimos. Até 3 salários mínimos. Até 4 salários mínimos. Até 5 salários mínimos. Até 6 salários mínimos. Até 7 salários mínimos. Até 8 salários mínimos. Até 9 salários mínimos. Acima de 10 salários mínimos.
q22	Até que etapa de escolarização seu pai concluiu (ou da pessoa que o/a criou como pai)?	Não tive pai ou pessoa que exerceu tal papel na minha criação. Nenhuma. Ensino fundamental: 1º ao 5º ano. Ensino fundamental: 6º ao 9º ano. Ensino médio. Ensino superior.
q23	Até que etapa de escolarização sua mãe concluiu (ou da pessoa que o/a criou como mãe)?	Não tive mãe ou pessoa que exerceu tal papel na minha criação. Nenhuma. Ensino fundamental: 1º ao 5º ano. Ensino fundamental: 6º ao 9º ano. Ensino médio. Ensino superior.
q25	Você possui acesso à internet?	Sim Não
q27	Qual ensino fundamental você cursou?	Ensino fundamental em idade própria. Educação de Jovens e Adultos (EJA). Supletivo. Outros.
q29	Em que tipo de escola você cursou o ensino fundamental?	Somente em escola pública. Maior parte em escola pública. Maior parte em escola particular sem bolsa. Maior parte em escola particular com bolsa. Somente em escola particular sem bolsa. Somente em escola particular com bolsa.
q45	A renda mensal per capita de sua família em SM (Salário Mínimo):	Até 0,5 SM. Acima de 0,5 até 1 SM. Acima de 1 até 1,5 SM. Acima de 1,5 até 2,5 SM. Acima de 2,5 até 3 SM. Acima de 3 SM.

Fonte: O Autor.

Nessa tarefa foram selecionadas inscrições que geraram uma matrícula na instituição após a aprovação do candidato. Isso foi necessário, pois nem sempre que um candidato é aprovado o mesmo efetua sua matrícula. Para esta tarefa houve a necessidade de consultar o registro acadêmico do IFPE. Assim, cada registro da tabela de inscrição foi analisado para saber se de fato ela originou uma matrícula.

A consulta no registro acadêmico foi realizada através de uma API (*Application Programming Interface*), disponibilizada pela DADT. Para realizar a conexão com API, criamos um script em Python, na qual ele faz a leitura de cada registro da tabela inscricao, passando o número da inscricao para a chamada da API, caso seja encontrado uma matrícula para esse número de inscrição o script inclui esse registro em um novo banco de dados. Esta consulta teve como data base a situação do estudante no mês de Março de 2020.

Esta consulta também serviu para definir a situação em que se encontra a matrícula que foi gerada. Foi criado um novo atributo na tabela ***inscricao*** chamado de **SituacaoStatus**, onde 0 indica estudantes que permaneceram ou permanecem em seu curso e 1 para aqueles que se evadiram. Consideramos evadidos aqueles estudantes que possuíam o status de evadido no registro acadêmico, o que acontece quando ele atinge 25% de faltas durante algum período letivo e se torna automaticamente evadido. **SituacaoStatus** será a variável alvo para os nossos modelos de classificação.

Após essa análise, chegamos à seguinte quantidade de candidatos que efetuaram a matrícula na modalidade subsequente conforme a tabela 5. A tabela também indica o número de estudantes conforme sua situação.

Tabela 5 – Matrículas da modalidade subsequente

Ano de Ingresso	Nº de Matrículas Efetuadas	Nº de Estudantes em situação regular	Nº de Estudantes Evadidos
2016.1	1.102	573	529
2016.2	54	29	25
2017.1	1.525	839	686
2017.2	1.143	668	475
2018.1	1.312	800	512
2018.2	1.416	974	442
2019.1	1.629	1.189	440
2019.2	1.571	1.336	235
Total	9.752	6.408	3.344

Fonte: O Autor

4.3.2 Integração e Formatação dos dados

Nesta tarefa buscou-se combinar as múltiplas bases de dados de todos os anos e concatenar as tabelas ***inscricao*** com sua tabela ***resposta_questionario*** correspondente afim de facilicar a compreensão e o desepenho dos classificadores. A tabela 6 mostra a base de dados final que foi utilizada nos algoritmos de AM.

Tabela 6 – Dataset utilizado nos modelo de classificação

ID	Atributo	Descrição	Tipo
1	Cor	Indica a cor do candidato	Categórica
2	Sexo	Indica o sexo do candidato	Categórica
3	EstadoCivil	Indica o estado civil do candidato	Categórica
4	Cotista	Indica se o candidato optou pelo sistema cotas	Categórica
5	Portugues	Nota da prova de português	Numérica
6	Matematica	Nota da prova de matemática	Numérica
7	Consgerais	Nota da prova de conhecimentos gerais	Numérica
8	q8	Pergunta q8 do questionário	Categórica
9	q10	Pergunta q10 do questionário	Categórica
10	q15	Pergunta q15 do questionário	Categórica
11	q18	Pergunta q18 do questionário	Categórica
12	q19	Pergunta q19 do questionário	Categórica
13	q22	Pergunta q22 do questionário	Categórica
14	q23	Pergunta q23 do questionário	Categórica
15	q25	Pergunta q25 do questionário	Categórica
16	q27	Pergunta q27 do questionário	Categórica
17	q29	Pergunta q28 do questionário	Categórica
18	q45	Pergunta q45 do questionário	Categórica
19	SituacaoStatus	Indica se o estudante evadiu ou não	Categórica

Fonte: O Autor

4.4 Modelagem

Primeiramente, nesta fase selecionamos qual a técnica de AM a ser utilizada. Para esse estudo, foram utilizados os algoritmos supervisionados para construção de um modelo preditivo de classificação binária com base nos atributos da Tabela X. Nossa

classe alvo é o atributo **SituacaoAluno**, que indica se o estudante abandonou o seu curso ou não.

Os Algoritmos supervisionados de classificação utilizados foram:

- **DecisionTreeClassifier**
- **RandomForestClassifier**
- **XGBClassifier**

Os algoritmos *DecisionTreeClassifier* e *RandomForestClassifier* estão disponíveis na biblioteca sklearn[3]. Os mesmos foram escolhidos por já estarem presentes em grande parte dos trabalhos relacionados. Já o *XGBClassifier* pode ser encontrado em seu site oficial[4] e teve critério de escolha definido pelo bom histórico em problemas de classificação e não sendo encontrado em outros trabalhos que envolvam predição de evasão.

Para cada algoritmo acima, foi aplicada a técnica de validação cruzada com 5 subconjuntos (*5-fold cross validation*).

A validação cruzada é uma técnica utilizada para avaliação e desempenho de modelos de AM. Na validação cruzada, o conjunto de dados é particionado em subconjuntos (*folds*) mutuamente exclusivos, um elemento do conjunto é separado para o conjunto de teste e os demais utilizados para o conjunto de treinamento. Esse processo é repetido *k* vezes, alternando-se os elementos dos grupos de teste e treinamento (SOARES, 2020).

A base de dados utilizada foi dividida em uma base para treino e uma base para teste. A base treino contou com os conjuntos de dados dos anos de 2016, 2017 e 2018. Totalizando 3.883 estudantes não evadidos e 2.669 estudantes em situação de evasão. A base para teste foi composta apenas pelo ano de 2019, pois os candidatos deste ano ingressaram no IFPE respectivamente nos meses de fevereiro e agosto. Assim, na data da consulta ao registro acadêmico (Março/2020) essa base contava com 2.525 estudantes em situação regular e 675 estudantes já evadidos.

4.5 Avaliação

Nesta etapa definimos a forma em que os modelos de AM empregados serão avaliados.

[3] <https://scikit-learn.org/stable/>
[4] <https://xgboost.readthedocs.io/>

Para analisar os resultados dos nossos modelos, utilizamos a matriz de confusão para medir a performance dos algoritmos de classificação. A matriz é representada por uma tabela com 4 diferentes combinações:

- **Verdadeiro Positivo (VP)**: Indica casos em que o estudante que não evadiu do curso e foi classificado corretamente como ”não evadido”.
- **Verdadeiro Negativo (VN)**: Representa casos em que o estudante evadiu e em que o modelo previu corretamente que ele de fato ”evadiu”.
- **Falso Positivo (FP)**: Indica casos em que o estudante que não evadiu do curso e foi classificado como ”evadido”.
- **Falso Negativo (FN)**: Indica casos em que o estudante que evadiu do curso e foi classificado como ”não evadido”.

A partir da matriz de confusão foram calculados 4 métricas de avaliação: **Acurácia**, **Precisão**, **Revocação** e **F1-Score**.

- **Acurácia**: indica o quanto dos nossos dados foram classificados corretamente. A = (VP + VN)/(VP + VN + FP + FN).
- **Precisão**: determina o valor de instâncias que foram classificadas com a classe ”1”(Evadiu) P = VP/(VP + FP).
- **Revocação**: indica o quanto frequentemente a classe positiva foi prevista R = VP/(VP + FN).
- **F1-Score**: apresenta a média harmônica entre Precisão e Revocação. No geral indica a precisão do modelo. F1 = 2 * (Precisão * Revocação)/(Precisão + Revocação).

Outra métrica de avaliação utilizada é a AUC (*Area Under the Curve*). A AUC representa o grau de separabilidade de um modelo. Um modelo com um alto grau de AUC significa que ele prediz melhores VP e VN. Sendo assim a métrica AUC pode ser interpretada como a probabilidade daquele modelo classificar uma instancia aleatória na classe positiva do que classificá-la na classe negativa. A área da AUC é calculada a partir da curva ROC (*Receiver Operating Characteristics*), que é um gráfico que exibe a performance de um modelo de classificação em todos os limites possíveis. Um bom classificador é avaliado quando sua curva esta mais próxima do ponto 1 no eito y, como demostra a figura.

5 Resultados

Nesta seção serão apresentados os resultados dos experimentos realizados de acordo com a metodologia adotada.

Na Tabela 7 são apresentados os resultados experimentais dos modelos preditivos para o conjunto de dados utilizado de acordo com as métricas utilizadas. Pode-se observar que o algoritmo *Random Forest* obteve a melhor acurácia com 64%, seguida do algoritmo *XGBoost*

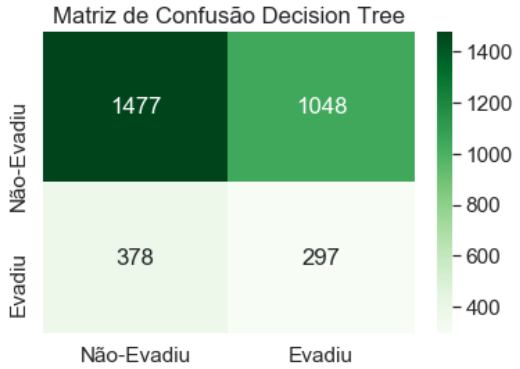
Tabela 7 – Desempenho dos algoritmos com validação cruzada 5-*Folds*

	Decision Tree	Random Forest	XGBoost
Acurácia	0.551251	**0.627541**	0.3890625
Precisão	0.503371	0.5015896	**0.515000**
Revocação	0.504921	0.5018320	**0.524272**
F1 Score	0.504144	0.501710	**0.519594**

Fonte: O Autor

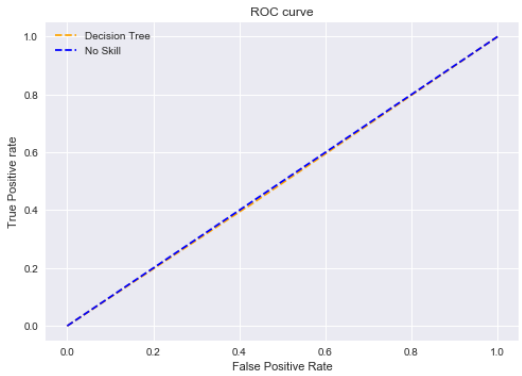
5.1 Algoritmo Decision Tree

Figura 6 – Matriz de Confusão Decision Tree



Fonte: O Autor

Figura 7 – Algoritmo Decision Tree

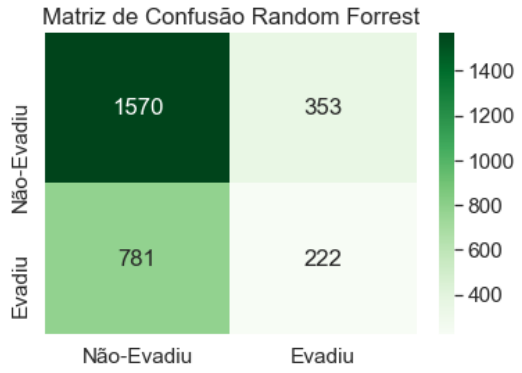


Fonte: O Autor

As figuras 6 e 7 demonstram o resultado deste algoritmo. A AUC atingiu apenas 0.49661. Conforme indica a Matriz de Confusão deste modelo na figura 6, o algoritmo classificou corretamente apenas 297 registros de estudantes evadidos e 1.477 para não evadidos. Seu ponto negativo fica com os 1.048 registros que foram classificados como evadidos mas estão em situação regular.

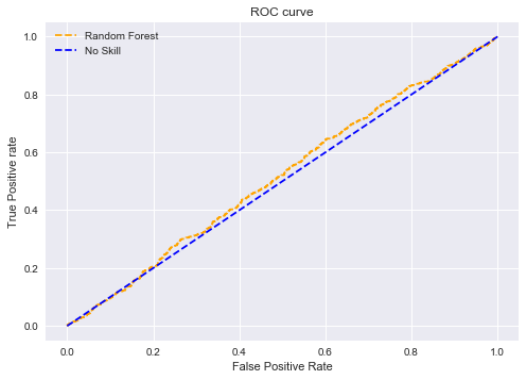
5.2 Algoritmo Random Forest

Figura 8 – Matriz de Confusão Random Forrest



Fonte: O Autor.

Figura 9 – Algoritmo Random Forrest

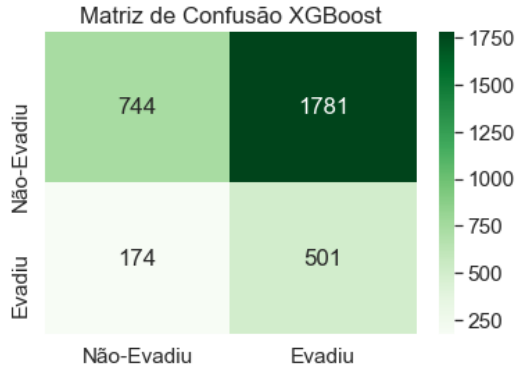


Fonte: O Autor.

As figuras 8 e 9 apresentam o resultado deste algoritmo. A AUC apresentou um valor de 0.51658. Este algoritmo indicou um valor baixo para sua Precisão. Apesar de sua Acurácia apresentar um valor de 0,6275, sendo considerada a mais alta entre os outros modelos.

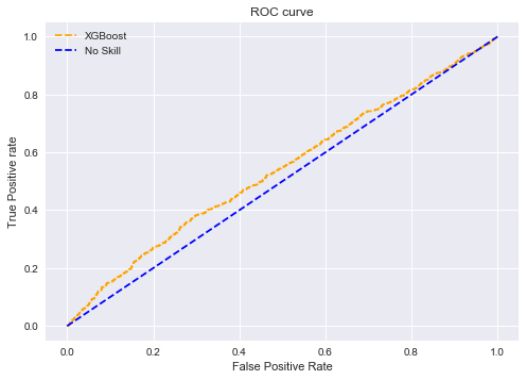
5.3 Algoritmo XGBoost

Figura 10 – Matriz de Confusão XGBoost



Fonte: O Autor.

Figura 11 – Algoritmo XGBoost



Fonte: O Autor.

As figuras 10 e 11 apresentam o resultado deste algoritmo. Conforme a matriz de confusão na figura 10 este algoritmo foi o melhor que classificou corretamente estudantes evadidos, 501 registros. Atingindo uma proporção de mais de 74% de acerto. Assim, foi o algoritmo que obteve o melhor valor de revocação. Um valor considerado bom para prever potenciais candidatos a evadirem. Por outro lado, o algoritmo apresentou um elevado número de registros com o valor Falso Positivo, indicando que 1.781 registros se evadiram mas se encontram matriculados normalmente. A AUC apresentou um valor de 0.54020.

6 Conclusão

Considerando os anos recentes, a quantidade de evasões nas instituições de ensino aumentam mais a cada dia. Afetando não apenas a instituição em si como também o futuro daquele estudante. Combater a evasão escolar não é uma tarefa fácil, em muitos casos ela depende de fatores que são intangíveis para o estudante.

Detectar um perfil evasor a partir do momento em que o candidato se inscreve no processo de vestibular envolve uma série de fatores pessoais e socioeconômicos, muito mais do que fatores exclusivamente acadêmicos. Diante dos resultados apresentados, nota-se que nenhum dos três modelos comparados atente de forma satisfatória detectar se aquele candidato irá se evadir ou não após sua aprovação no vestibular. Os mesmos apresentam uma alta taxa de falsos negativos o que pode levar a uma tomada de decisão incorreta e imprecisa. Então, pode-se afirmar que os atributos selecionados não são totalmente relevantes para detectar uma possível evasão no momento em que o estudante ingressar na instituição de ensino.

Assim, este trabalho procurou trazer uma nova opção para se combater à evasão. É notório que a Mineração de Dados Educacionais vem crescendo nos últimos anos e gerando conhecimento através de dados antes inexplorados. Este estudo se torna inovador pelo fato de ter trabalhado com dados de estudantes antes do seu início na vida acadêmica, o que não foi encontrado até então na literatura.

Ademais, diante dessas considerações podemos afirmar que este trabalho não obteve resultados fortemente expressivos a ponto de se tornar uma ferramenta que auxilie a instituição de ensino no momento em que ela recebe seus futuros estudantes.

6.1 Trabalhos Futuros

Para trabalhos futuros, iremos aprimorar a seleção de atributos que possam influenciar na predição de potenciais evasores, visto que os atributos utilizados neste trabalho não foram suficientes para obter uma alta taxa de acerto. Alguns atributos que podem contribuir para a classificação são por exemplo notas das disciplinas cursadas, assim como a quantidade de faltas do estudante, indicadores escolares todos eles referentes onde o estudante cursou seu ensino médio. Também pretendemos desenvolver uma uma aplicação e implantar ela nos próximos processos seletivos do IFPE. Além de utilizar outros algoritmos de classificação e comparar com os resultados atuais.

Referências

ALBAN, M.; MAURICIO, D. Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology*, v. 12, n. 4, p. 1–12, 2019. Citado na página 12.

ALVAREZ, K. R.; MATOS, R. P. Permanência e êxito escolar nos institutos federais. *Ensino em Foco*, v. 3, n. 6, p. 106–115, 2020. Citado na página 12.

ARRUDA, D. Z. M. Evasão escolar no ensino técnico: um estudo de caso numa escola técnica do centro paula souza. Universidade Estadual Paulista (UNESP), 2019. Citado na página 10.

ASIF, R. et al. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, Elsevier, v. 113, p. 177–194, 2017. Citado na página 15.

AULCK, L. et al. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016. Citado 3 vezes nas páginas 15, 21 e 23.

BAKER, R. et al. Data mining for education. *International encyclopedia of education*, Elsevier Oxford, UK, v. 7, n. 3, p. 112–118, 2010. Citado na página 16.

BARBOSA, A.; SANTOS, E.; PORDEUS, J. P. A machine learning approach to identify and prioritize college students at risk of dropping out. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1497. Citado 2 vezes nas páginas 21 e 23.

BELTRAN, C. A. R. et al. Plataforma de aprendizado de maquina para detecção e monitoramento de alunos com risco de evasão. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2019. v. 30, n. 1, p. 1591. Citado 2 vezes nas páginas 22 e 23.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996. estabelece as diretrizes e bases da educação nacional. *Diário Oficial [da] República Federativa do Brasil*, 1996. Citado na página 10.

BRASIL, M. Comissão especial de estudos sobre a evasão nas universidades públicas brasileiras. *http://www. dominiopublico. gov. br/download/texto/me001613. pdf Acesso em*, v. 15, n. 01, p. 2007, 1997. Citado 2 vezes nas páginas 10 e 11.

CHAPMAN, P. et al. *CRISP-DM 1.0 Step-by-step data mining guide*. [S.l.], 2000. Disponível em: <https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf>. Citado na página 24.

COLPANI, R. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. *Informática na educação: teoria & prática*, v. 21, n. 3, 2018. Citado na página 10.

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, 2013. Citado na página 16.

FILHO, R. B. S.; ARAúJO, R. M. d. L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. *Educação Por Escrito*, v. 8, n. 1, p. 35–48, jun. 2017. Disponível em: <https://revistaseletronicas.pucrs.br/index.php/porescrito/article/view/24527>. Citado na página 14.

LÜSCHER, A. Z.; DORE, R. Política educacional no brasil: educação técnica e abandono escolar. *Revista Brasileira de Pós-Graduação*, v. 8, n. 1, 2011. Citado na página 14.

MANHÃES, L. M. B. et al. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In: *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)*. [S.l.: s.n.], 2012. v. 1, n. 1. Citado na página 15.

MARIA, W.; DAMIANI, J. L.; PEREIRA, M. Rede bayesiana para previsão de evasão escolar. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2016. v. 5, n. 1, p. 920. Citado 2 vezes nas páginas 21 e 23.

NARCISO, L. G. d. S. et al. Análise da evasão nos cursos técnicos do instituto federal do norte de minas gerais-campus arinos: exclusão da escola ou exclusão na escola? 2015. Citado na página 11.

OLIVEIRA, F. A. d. C. et al. Evasão escolar no ensino técnico profissionalizante: um estudo de caso no instituto federal goiano–campus ceres. Instituto Federal Goiano, 2019. Citado na página 10.

PEREZ, B.; CASTELLANOS, C.; CORREAL, D. Applying data mining techniques to predict student dropout: a case study. In: IEEE. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. [S.l.], 2018. p. 1–6. Citado na página 15.

PRIYAM, A. et al. Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, Citeseer, v. 3, n. 2, p. 334–337, 2013. Citado na página 18.

RESENDE, P. A. A.; DRUMMOND, A. C. A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 3, p. 1–36, 2018. Citado na página 18.

RIGO, S. J. et al. 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. 01, p. 132, 2014. Citado na página 16.

RODRIGUES, R. L. et al. A literatura brasileira sobre mineração de dados educacionais. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2014. v. 3, n. 1, p. 621. Citado na página 16.

ROLIM V.B, C. F. R. F. R. Reconhecimento de padrões aplicados a comentários de fóruns educacionais. 2014. Citado na página 18.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Ieee, v. 40, n. 6, p. 601–618, 2010. Citado na página 12.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013. Citado na página 15.

ROMERO, C. et al. *Handbook of educational data mining*. [S.l.]: CRC press, 2010. Citado na página 15.

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–e06911570, 2020. Citado na página 10.

SILVA, E. et al. Evasão no ensino básico da rede pública municipal de juiz de fora: uma análise com mineração de dados. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2019. v. 30, n. 1, p. 1371. Citado 2 vezes nas páginas 21 e 23.

SOARES, L. C. C. P. Soluções tecnológicas para o problema da evasão universitária, sob a óptica de ferramentas de inteligência artificial. UFT, 2020. Citado na página 34.

TASNIM, N.; PAUL, M. K.; SATTAR, A. S. Identification of drop out students using educational data mining. In: IEEE. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. [S.l.], 2019. p. 1–5. Citado 2 vezes nas páginas 22 e 23.

TÁVORA, L. et al. Institutos federais de educação, ciência e tecnologia e o apoio a inovação tecnológica: análises e recomendações. In: *XVI Congresso Latino-Iberoamericano de Gestão da Tecnologia–ALTEC*. [S.l.: s.n.], 2015. Citado na página 10.