Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática

Daniel Victor Saraiva[1], **Silas Santiago Lopes Pereira**[1], **Erica de Lima Gallindo**[1], **Reinaldo Bezerra Braga**[1], **Carina Teixeira de Oliveira**[1]

[1] Laboratório de Redes de Computadores e Sistemas (LAR)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Caixa Postal 62.800-000 – Aracati – CE – Brasil

Abstract. *The high rate of dropout in technology courses is a challenge faced by a large number of educational institutions. Therefore, identifying the risk of student dropout is relevant once actions can be taken to encourage permanence and success. In this context, this paper presents a proposal for predicting dropouts at educational institutions considering academic and socioeconomic information of students. In particular, one of the proposal steps involves the application of Machine Learning algorithms. A case study is presented making use of a 8-years dataset of a Computer Technician course at IFCE. The results demonstrate the effectiveness of the proposal, reaching an accuracy about 97.97% to predict students' situation.*

Resumo. *O alto índice de evasão em cursos de tecnologia é um desafio enfrentado por muitas instituições de ensino. Por isso, identificar os estudantes em situação de risco de evasão é relevante para que possam ser realizadas ações de incentivo à permanência e êxito. Este trabalho apresenta uma proposta para predição de evasão considerando informações acadêmicas e socioeconômicas de estudantes. Em particular, uma das etapas da proposta engloba a aplicação de algoritmos de Aprendizagem de Máquina. Um estudo de caso é apresentado considerando uma base de dados de 8 anos de um curso Técnico em Informática do IFCE. Os resultados demonstram a eficácia da proposta, alcançando 97,97% de taxa de acerto na previsão da situação dos estudantes.*

1. Introdução

A educação, sobretudo a institucionalizada, permite que os seres humanos adquiram os conhecimentos e habilidades necessários para atuar na sociedade. Entretanto, mesmo diante das previsões constitucionais e legais [BRASIL 1988, BRASIL 1996], que estabelecem meios para o acesso à educação, permanência e êxito estudantil, a evasão na educação emerge como problema real. Nesse contexto, a evasão preocupa muitos profissionais e instituições de ensino, desde a educação básica até o nível superior de instituições públicas e privadas nas diferentes áreas de formação [BRASIL 2014, PNUD 2015, INEP 2017].

Diante dessa problemática, a previsão do desempenho acadêmico de estudantes é um desafio de pesquisa de grande importância, pois tanto os estudantes quanto as instituições podem se beneficiar dos resultados. Em outras palavras, as instituições podem melhorar a qualidade acadêmica e otimizar os recursos disponíveis para ajudar seus estudantes a concluírem seus estudos com êxito [Lei and Li 2015]. Enquanto isso, para

os gestores educacionais e professores, os resultados das pesquisas podem ser usados na tomada de ações alternativas para que os estudantes possam ter melhores resultados, permitindo assim, um desempenho acadêmico bem-sucedido [Lei and Li 2015].

A avaliação da aprendizagem é parte integrante do processo didático de qualquer instituição de ensino, seja esta pública ou privada ou do nível de ensino oferecido [Correia et al. 2016]. Conforme [Depresbiteris 1998], planejamento e avaliação são tarefas indissociáveis, de modo que constituem um processo único para os quais devem ser definidos objetivos, conteúdos e estratégias de ensino, critérios e modos de avaliar. O conhecimento antecipado da situação do estudante constitui informação valiosa para a tomada de decisão do núcleo gestor da instituição no que se refere as políticas para combate à evasão e retenção, de maneira a permitir uma avaliação antecipada a nível institucional e de curso.

Nesse cenário, as pesquisas para previsão de desempenho de estudantes são geralmente baseadas na extração de conhecimento em bases de dados educacionais. Entretanto, analisar uma enorme quantidade de dados para encontrar informações úteis de forma resumida é uma tarefa custosa para o ser humano. Desse modo, técnicas de Mineração de Dados usando métodos confiáveis de Aprendizado de Máquina - ou seja, métodos matemáticos para treinar algoritmos - podem ser utilizadas no campo educacional para ampliar a compreensão do processo de aprendizagem, identificando e avaliando as variáveis que interferem no desempenho dos estudantes [Zhang and Li 2018, Hegde and Prageeth 2018].

O Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) oferta cursos que abrangem o ensino básico, técnico, graduação e pós-graduação por meio da tríade Ensino, Pesquisa e Extensão. Segundo a própria instituição, os seus índices de evasão têm sido significativos, contrariando a perspectiva de universalização do acesso à educação e da garantia da permanência [IFCE 2017]. Por exemplo, entre o semestre 2009/1 até outubro de 2018 [IFCE 2018] foram realizadas 6.244 matrículas em cursos de Técnico em Informática no IFCE. Do total de ingressantes, 3.091 (49,50%) dos estudantes saíram de seus cursos antes de concluí-los, ou seja, a taxa de evasão está próxima de representar a metade dos ingressantes, um número preocupante para a instituição. Neste grupo, estão incluídos aqueles ingressantes cujas matrículas estão atualmente nas seguintes situações: abandono, transferência (interna ou externa), cancelada (compulsória ou voluntária) e falecimento[1]. A primeira categoria é denominada *egressos sem êxito*.

Além destes, 1.296 (20,76%) ingressantes concluíram seus cursos (egressos com êxito) e 1.667 (26,70%) estão vinculados ao curso, ou seja, estão frequentando o curso (em curso). Para os estudantes que encontram-se com os estudos interrompidos (trancado ou em intercâmbio) esse valor é de 133 (2,13%). Já para os que encontram-se com suas matrículas integralizadas em fase escolar, representando que o estudante já integralizou a carga horária de disciplinas obrigatórias mas depende de outras etapas para a obtenção da certificação (ex: aguardando cumprir carga horária de estágio), esse valor é de 57 (0,91%).

Diante disso, tanto no IFCE quanto em outras instituições de ensino, a evasão dos estudantes representa um grande desafio. Consequentemente, existe a necessidade contínua do desenvolvimento de soluções que apontem caminhos, métodos e ferramentas

[1]O total de falecidos até outubro de 2018 é de 1 ingressante.

que auxiliem as instituições no enfrentamento desse problema.

Neste contexto, este trabalho apresenta uma proposta para predição de evasão considerando informações acadêmicas e socioeconômicas de estudantes de um curso de Técnico em Informática do IFCE. Uma das etapas da proposta engloba a aplicação de Mineração de Dados utilizando algoritmos de Aprendizagem de Máquina para detecção precoce do risco de evasão dos estudantes. Com os resultados das previsões, o objetivo é que gestores e professores possam realizar melhorias nas metodologias de ensino-aprendizagem, nos processos avaliativos e nas ações de incentivo à permanência e êxito de estudantes que estão com maior risco de evasão.

O restante do trabalho está organizado da seguinte maneira: a Seção 2 apresenta os principais conceitos utilizados nesta pesquisa relacionados à mineração de dados e descoberta de conhecimentos; a Seção 3 detalha alguns trabalhos relacionados na área de mineração de dados educacionais; a Seção 4 apresenta a proposta desta pesquisa para previsão da situação de estudantes; a Seção 5 apresenta e discute os resultados da proposta; por fim, a Seção 6 apresenta as considerações finais e direcionamentos para trabalhos futuros.

2. Mineração de Dados e Descoberta de Conhecimentos

Com o crescimento exponencial da quantidade de dados armazenados e considerando que os dados não podem ser analisados manualmente em virtude do grande volume de registros, é necessário um trabalho de busca detalhado com o objetivo de descobrir conhecimento que pode estar implícito [De Castro and Ferrari 2016, Silva et al. 2016]. Nesse contexto, essa busca deve estar associada a um processo analítico, sistemático e, até onde possível, automatizado [Silva et al. 2016]. Conforme destacado no trabalho de [Silva et al. 2016], é nesse cenário de descoberta de conhecimento em base de dados que a Mineração de Dados, do inglês *Data Mining*, é definida.

Mineração de dados é um processo automático ou semiautomático de exploração para descoberta de padrões relevantes em bases de dados [Silva et al. 2016]. A partir da descoberta de padrões, é possível gerar conhecimento útil para um processo de tomada de decisão, ou seja, algo anteriormente não conhecido e que tenha valor para o domínio em que será aplicado [Faceli et al. 2011, Silva et al. 2016]. Para isso, são necessárias técnicas implementadas por meio de algoritmos computacionais capazes de receber, como entrada, uma base de dados de fatos ocorridos no mundo real e devolver, como saída, um padrão de comportamento que possa ser expresso de diferentes formas, por exemplo, como uma regra de associação, uma função de mapeamento ou uma modelagem de um perfil [Silva et al. 2016]. Segundo Faceli *et al.* [Faceli et al. 2011], técnicas de Aprendizagem de Máquinas (AM) estão entre as mais empregadas no processo de mineração de dados, e diversas aplicações de sucesso de AM em mineração de dados são continuamente mencionadas.

Aprendizagem de máquina é uma área de pesquisa que se preocupa em desenvolver programas computacionais capazes de melhorar seus desempenhos automaticamente por meio da experiência [Mitchell 1997]. Para Faceli *et al.* [Faceli et al. 2011], métodos de aprendizagem de máquina devem ser capazes de criar, a partir da experiência passada, uma hipótese ou função na qual se obtém conclusões genéricas a partir de um conjunto de dados que representa o problema a ser resolvido.

A aplicação de métodos de mineração de dados e aprendizagem de máquina na educação tem sido vista como um campo interdisciplinar emergente. Essa nova área de pesquisa é chamada de Mineração de Dados Educacionais (*Educational Data Mining* - EDM) [Devasia et al. 2016, Marwaha and Ahuja 2017, Hegde and Prageeth 2018]. A EDM é usada para estudar os dados disponíveis no contexto educacional e extrair valor das informações ocultas em bases de dados de instituições de ensino. Essas informações podem ser usadas em vários processos educacionais, como previsão de matrículas em cursos, estimativa da taxa de evasão de estudantes e previsão de desempenho acadêmico [Tekin 2014, Yukselturk et al. 2014].

3. Trabalhos Relacionados

Esta seção apresenta uma revisão bibliográfica de pesquisas presentes na literatura que realizaram previsão de risco de evasão de estudantes. Em particular, é dado um foco especial aos trabalhos publicados entre os anos de 2016 e 2018. Dentre tais trabalhos, quatro pesquisas usam bases de dados de instituições de ensino do Brasil e outros quatro trabalhos destacam pesquisas realizadas em outros países.

3.1. Cenário Nacional

No trabalho de [Maria et al. 2016], os autores apresentam um sistema computacional que utiliza redes bayesianas para obtenção da probabilidade de evasão de estudantes. A solução permite que um gestor simule os possíveis cenários para os estudantes, a fim de minimizar as chances de evasão. A predição é realizada com base nas características de dados de estudantes dos cursos técnicos coletadas no Sistema de Gestão de Negócios (SGN) utilizado pelo Serviço Nacional de Aprendizagem Industrial (SENAI) na unidade de Tubarão - SC. Para a construção do modelo, é realizado um levantamento das principais informações que podem ser relacionadas aos estudantes para posterior predição da evasão. A base de dados utilizada no trabalho possui 666 estudantes. Desse total, 337 (50,6%) evadiram dos cursos e 329 (49,4%) não evadiram. O resultado final do desempenho da rede bayesiana modelada é de 85,6% de taxa de acerto.

Paz e Cazella [Paz and Cazella 2017] apresentam os resultados de um estudo que busca identificar perfis de estudantes com potencial de evasão em cursos de graduação de uma Universidade Comunitária[2] do Rio Grande do Sul. São utilizados dados amostrais de 4.601 instâncias e 7 atributos de estudantes matriculados em cursos de graduação do semestre 2016/2 de todos os campi da Universidade. O algoritmo J48 [Quinlan 1993] é utilizado na etapa de predição. Esse algoritmo permite visualizar o modelo preditivo em um formato de árvore de decisão. Os experimentos realizados têm acurácias superiores a 91%.

O trabalho realizado por [Lanes and Alcântara 2018] tem como objetivo identificar o subconjunto dos estudantes de graduação da Universidade Federal do Rio Grande (FURG) que apresentam risco de evasão. Para isso, são analisadas diferentes informações dos estudantes que evadiram ou concluíram seus cursos de graduação entre 2012 e 2017. A base de dados utilizada no trabalho é coletada do sistema acadêmico da FURG, contendo informações sobre estudantes de 12 cursos de graduação de diferentes áreas do

[2]As universidades comunitárias constituem um segmento de Instituição de Ensino Superior - IES, cujos fins estão voltados, além da educação, aos serviços sociais e à comunidade. Mesmo a origem de seus recursos sendo oriunda de mensalidades, não apresenta fins lucrativos [Veiga et al. 2012]

conhecimento. 916 registros de estudantes são utilizados, dentre os quais, 720 (78,60%) pertencem à classe denominada evadido e 196 (21,40%) à classe denominada concluinte. A classificação é realizada utilizando o algoritmo J48 [Quinlan 1993] para processar o *dataset* e gerar uma árvore de decisão. De acordo com os resultados experimentais, o algoritmo J48 apresenta acurácia de 90,7%.

Em [Gonçalves et al. 2018], os autores realizam uma pesquisa que tem como objetivo extrair conhecimento útil de dados sobre os estudantes de graduação do Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA). A proposta é tentar entender os problemas da evasão do referido instituto. No artigo, são usados três algoritmos: *Naive Bayes*, Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) e J48. Além disso, três abordagens de seleção de atributos são também testadas: manual, seleção baseada em correlação e ganho de informação. A base de dados do trabalho é definida com 40 atributos e 574 instâncias, sendo 287 estudantes que concluíram os cursos e 287 que evadiram. Testes indicaram que os melhores resultados são obtidos pelo J48 com taxa de acerto de 98.08% por meio da seleção baseada em correlação.

3.2. Cenário Internacional

No trabalho [Hegde and Prageeth 2018], os autores apresentam uma metodologia para prever a evasão de estudantes usando o algoritmo de classificação *Naive Bayes*. Os dados são coletados por meio de um questionário com perguntas direcionadas aos estudantes de primeiro semestre e para estudantes de terceiro semestre. O objetivo de utilizar dois questionários é distinguir informações de um estudante que entra e de um estudante que já está em uma instituição de ensino. O conjunto de dados utilizado no experimento possui informações de 50 estudantes e 24 atributos diferentes. A taxa de acerto do desempenho do classificador *Naive Bayes* é de 72%.

Solis *et al.* [Solis et al. 2018] utilizam algoritmos de aprendizado de máquina para prever desistências de estudantes universitários em um programa de graduação do Instituto Tecnológico da Costa Rica (ITCR). A amostra é composta por estudantes matriculados entre os anos de 2011 e 2016. São utilizadas informações de 15.720 estudantes. A capacidade de generalização de quatro algoritmos para prever o abandono são avaliadas: *Random Forest*, Perceptron Multicamadas (*Multilayer Perceptron* - MLP), Máquinas de Vetores de Suporte (SVM) e Regressão Logística. Após analisar os resultados, o melhor algoritmo para classificar as desistências é o *Random Forest* com *mtry* = 10 (número de variáveis aleatoriamente amostradas como candidatas em cada divisão) que obteve 85% de precisão.

Um estudo de caso concentrado na detecção de abandonos de estudantes de um curso de graduação em Engenharia de Sistemas (SE) em uma universidade privada em Bogotá, Colômbia é apresentado em [Perez et al. 2018]. No trabalho é realizada uma comparação entre os modelos de Árvore de Decisão, Regressão Logística e *Naive Bayes*. O conjunto de dados usado no trabalho possui informações de 802 estudantes matriculados no Programa de Ciência da Computação da universidade. O foco do trabalho está nos estudantes que ingressaram na universidade do primeiro semestre de 2004 até o segundo semestre de 2010. A árvore de decisão alcançou o melhor desempenho com 94% de acurácia, seguida por Regressão Logística (92%) e *Naive Bayes* (87%).

Segundo Dharmawan *et al.* [Dharmawan et al. 2018], pesquisas para prevenir

evasão de estudantes se concentram em fatores acadêmicos como determinantes para desistências. No entanto, há casos de estudantes que desistem de seus cursos e a causa não está relacionada à fatores acadêmicos. Na visão dos autores, isso levanta a hipótese de que os potenciais abandonos podem ser determinados a partir de fatores não acadêmicos. Para eles, há cinco critérios não acadêmicos que podem ser usados para prevenir abandonos: dados demográficos, interação social, finanças, motivação e fatores pessoais. Nesse sentido, o estudo analisa fatores não acadêmicos que podem influenciar no desempenho dos estudantes. Para realizar a previsão da evasão, são utilizados três métodos de classificação: Árvore de Decisão, SVM e K-vizinhos mais próximos (K-NN). A base de dados gerada no trabalho possui informações de 103 estudantes. Com base nos resultados dos experimentos, os algoritmos de Árvore de Decisão e SVM tem nível de precisão de 66%.

3.3. Comparativo da Proposta com os Trabalhos Relacionados

Os trabalhos apresentados nos cenários nacional e internacional realizam a predição da evasão de estudantes em diferentes situações como, por exemplo, diferentes níveis de ensino (graduação e técnico). Nesse contexto, diferente dos trabalhos apresentados, a proposta deste artigo não analisa apenas a saída do estudante por nível de ensino ou saída do estudante de um curso. Aqui, a evasão é analisada por nível de ensino, curso, campus e modalidade por considerar que são granularidades que se aproximam das causas da evasão, proporcionando uma previsão mais aprofundada e eficiente da situação dos estudantes.

Uma outra vantagem dessa proposta em relação aos trabalhos relacionados é a utilização e comparação de técnicas de aprendizagem de máquinas baseadas nos paradigmas de aprendizagem mais utilizados na literatura para escolha do melhor algoritmo na previsão da situação de estudantes: probabilístico (*Naive Bayes*), baseados em distância (K-Vizinhos mais Próximos - KNN), baseados em procura (Árvore de Decisão e *Random Forest*) e baseados em otimização (Redes Neurais Artificiais e Máquinas de Vetores de Suporte - SVM).

4. Proposta

A mineração de dados converte dados brutos de sistemas educacionais em conhecimento que pode ser usado por desenvolvedores de software educacional, professores, pesquisadores educacionais, entre outros. Nesse contexto, as etapas da mineração precisam acontecer de maneira efetiva e consistente, de forma que possibilitem resultados adequados que auxiliem na tomada de decisões. Assim, os componentes e as etapas utilizadas nesse trabalho para descoberta de conhecimento em base de dados educacionais são apresentados na Figura 1.

4.1. Obtenção dos Dados

A primeira etapa do processo de descoberta de conhecimento consiste na obtenção de dados referentes ao domínio de estudo. Para o desenvolvimento desta pesquisa, a fonte primária de dados utilizada é de origem do banco de dados do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE). Os dados foram disponibilizados em arquivo no formato *Comma-separated values* (CSV) com informações acadêmicas e socioeconômicas dos estudantes da instituição.

Figura 1. Etapas do processo de descoberta de conhecimento [Fonte: Autores].

A base de dados explorada neste trabalho é referente aos ingressantes de todos os cursos de Técnico em Informática de 13 campi do IFCE dos últimos dez anos. Especificamente, entre os semestres 2009/1 e 2019/1. Os dados considerados são os que datam até o dia 08 de janeiro de 2019.

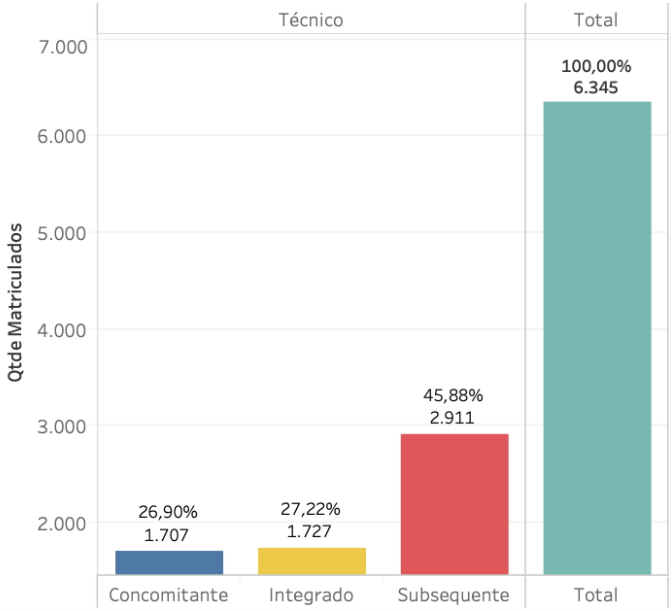

Figura 2. Quantidade de ingressantes em cursos de Técnico em Informática do IFCE entre os semestres 2009/1 e 2019/1 [Fonte: Autores].

Como ilustrado na Figura 2, essa base é composta por 6.345 ingressantes. Dentre esse total, 26,90% são ingressantes da modalidade de curso concomitante, 27,22% são ingressantes da modalidade integrado e 45,88% são da modalidade de curso subsequente. A modalidade concomitante é ofertada a quem ingressa no Ensino Médio ou já o esteja cursando, efetuando-se matrículas distintas para cada curso. A modalidade integrado é ofertada a quem já concluiu o Ensino Fundamental, com matrícula única na mesma instituição, promovendo habilitação profissional técnica de nível médio do estudante e conclusão da última etapa da educação básica. A modalidade subsequente é ofertada a quem já concluiu o Ensino Médio [BRASIL 2014].

4.2. Análise Visual dos Dados

A fase de análise visual dos dados é expressa como a segunda etapa da Figura 1. Ela engloba a compreensão dos dados de interesse a partir dos quais se deseja descobrir algum tipo de conhecimento.

O software *Tableau*[3] (versão 2018.2) é utilizado nessa etapa do processo para uma melhor compreensão dos dados existentes no arquivo disponibilizado. Pode-se definir o *Tableau* como uma solução para visualização de dados interativos focada em *business intelligence*. Ele permite a criação de filtros, imagens, painéis interativos, entre outras funcionalidades, que facilitam a exploração e análise dos dados.

A Figura 3 ilustra um exemplo do painel de trabalho disponibilizado pelo *Tableau* para interação com a base de dados do IFCE. No caso específico da figura, o painel apresenta o total de ingressantes por sexo, a estrutura etária (idade do estudante ao ingressar no IFCE) e etnia.

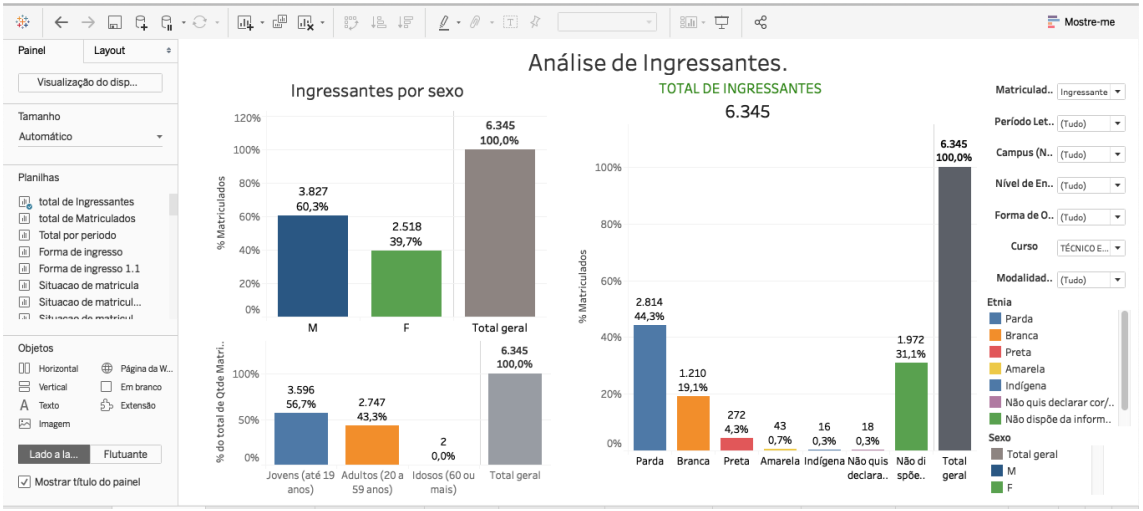

Figura 3. Exemplo de painel de trabalho do *Tableau* interagindo com a base de dados deste trabalho [Fonte: Autores].

A partir da análise visual dos dados do IFCE no *Tableau*, é possível obter uma série de análises úteis, como a quantidade e categorização dos perfis de estudantes evadidos da instituição (por idade, sexo, etnia, renda familiar, grau de instrução da mãe, etc). Ademais, é possível filtrar estudantes de acordo com o curso, situação de matrícula e outros diferentes atributos de interesse para a proposta.

4.3. Pré-processamento

Bases de dados podem apresentar diferentes características, dimensões e/ou formatos [Faceli et al. 2011]. Por exemplo, um conjunto de dados pode conter os seguintes tipos de valores: incorretos, inconsistentes, duplicados ou ausentes; os atributos podem ser independentes ou relacionados; os conjuntos de dados podem apresentar poucos ou muitos objetos que, por sua vez, podem ter um número pequeno ou elevado de atributos.

Dessa forma, após as etapas de obtenção e análise visual dos dados, técnicas de pré-processamento de dados são desejáveis para melhorar a qualidade dos dados por meio da eliminação ou minimização dos problemas mencionados. Consequentemente, após a aplicação de técnicas de preparação dos dados, os algoritmos de mineração são capazes de construir modelos mais fiéis à distribuição real dos dados [Faceli et al. 2011]. Algumas

[3]https://www.tableau.com

das principais tarefas do pré-processamento utilizadas neste trabalho são: eliminação manual de atributos; integração de dados; amostragem de dados, balanceamento do conjunto de dados; limpeza de dados; transformação de dados e redução de dimensionalidade.

4.4. Mineração de dados

A partir dos dados pré-processados, os conjuntos de dados são preparados para a criação de modelos. Os mesmos são submetidos ao processo de mineração de dados para extração de conhecimento relevante em relação à situação de evasão dos estudantes do IFCE.

Neste trabalho, são utilizados algoritmos com o objetivo de comparar os resultados de predição e, a partir dessa análise, definir qual o classificador a ser adotado nas demais fases da pesquisa. Para essa tarefa, são escolhidos no mínimo um algoritmo das principais categorias de classificadores: probabilístico - *Naive Bayes*; baseado em distância - *K-Nearest Neighbor* (KNN); baseados em procura - Árvore de Decisão e *Random Forest*; e baseados em otimização - Redes Neurais Artificiais (*Artificial Neural Networks* - ANNs) e Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM).

Essas escolhas foram tomadas após uma extensa revisão da literatura, tais como os trabalhos de [Mitchell 1997], [Duda and Hart 2001], [Witten and Frank 2005], [Wu et al. 2008] e [Faceli et al. 2011]; e os trabalhos apresentados na Seção 3, que apontam esses classificadores como os predominantes em pesquisas na área de mineração de dados educacionais.

4.5. Pós-processamento

A última etapa da Figura 1 é o pós-processamento. Nessa etapa, os resultados obtidos ou modelo são interpretados e usados para tomar decisões sobre o ambiente educacional. Nesta etapa, o conhecimento descoberto também é documentado para uso posterior. As técnicas de visualização também são muito úteis para mostrar os resultados de uma forma que seja mais fácil de interpretar os resultados. Em vez de mostrar apenas os resultados ou modelo obtido na mineração, nesta etapa, uma lista de sugestões ou conclusões sobre os resultados e como aplicá-los pode ser apresentada aos usuários.

5. Resultados e Discussões

Esta seção apresenta os principais resultados obtidos usando modelos preditivos para previsão da situação de estudantes em um curso de Técnico em Informática do IFCE, conforme o processo detalhado na Seção 4.

5.1. Obtenção dos Dados

Como definido anteriormente, este trabalho analisa a evasão dos estudantes do IFCE por nível de ensino, curso, campus e modalidade por considerar que são granularidades que podem permitir maior aproximação das causas da evasão. Nesse cenário, da base de dados representada na Figura 2, são selecionados dados de um curso Técnico em Informática do IFCE campus Aracati na modalidade concomitante. A base contém dados de 710 estudantes entre os semestres 2010/2 e 2018/2. A base possui quatro grupos de situações de matrículas. Considerando o objetivo deste trabalho em realizar classificação entre estudantes que concluíram e que evadiram do curso, são retirados da base os registros de matrículas de estudantes *em curso* e *com estudos interrompidos*. Além disso, são retiradas da base informações referentes a um estudante falecido. Ao final dessa etapa, a base é reduzida para 601 estudantes (230 egressos com êxito e 371 egressos sem êxito).

5.2. Análise Visual dos Dados

A Figura 4 mostra a situação da evasão do curso de Técnico em Informática por período letivo. Pode-se observar que em dez períodos, as porcentagens de egressos sem êxito passam dos 50% em relação ao total de ingressantes nesses períodos. No período 2015/1, por exemplo, 33 ingressantes evadiram do curso. Esse total representa 80% do total de ingressantes nesse período.

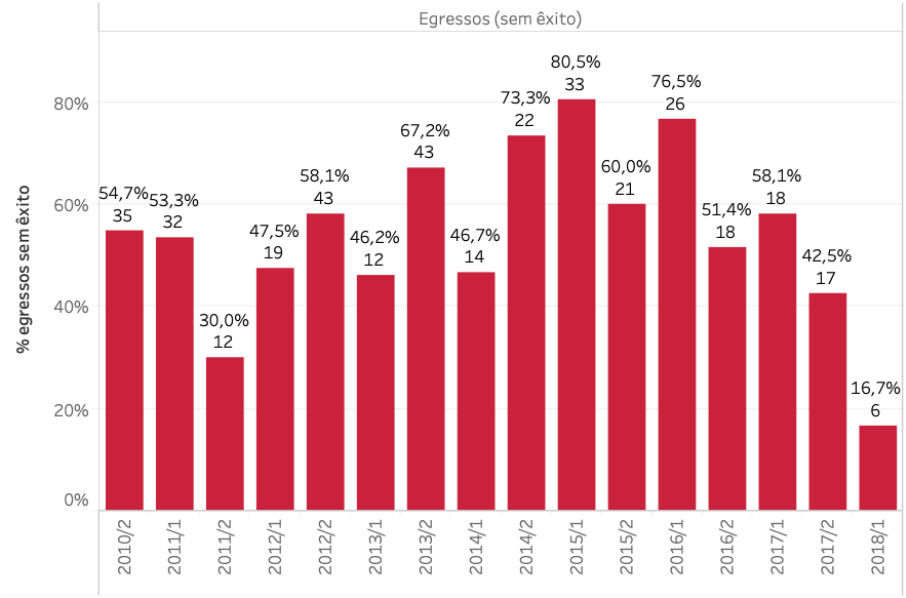

Figura 4. Egressos sem êxito do Técnico em Informática do campus Aracati por período [Fonte: Autores].

5.3. Pré-processamento

Para reduzir o desbalanceamento dos dados entre as duas classes (egressos com êxito e sem êxito) e para manter o conjunto de dados com boa quantidade de registros, são retiradas da base de dados 101 estudantes que evadiram e que possuem alta porcentagem de informações ausentes. Com isso, é gerada uma base de dados com 500 estudantes: 230 de egressos com êxito (46%) e 270 egressos sem êxito (54%).

Além disso, a base original contém informações sobre cada semestre cursado por um estudante ao longo de sua vida acadêmica. Assim, é realizada a transformação do parâmetro *situação de matrícula no semestre* em cinco novos parâmetros, que estão relacionados ao desempenho acadêmico do estudante: total de situações de períodos aprovados - situação que representa se o estudante foi aprovado em todas as disciplinas em que estava matriculado; total de situações de períodos aprovados parcialmente - situação que representa que o estudante ficou reprovado em, pelo menos, uma disciplina do período; total de situações de períodos aprovados com dependência - situação que representa que o estudante ficou reprovado em até duas disciplinas do período letivo - situação para os cursos técnicos; total de situações de períodos reprovados - situação que representa que o estudante foi reprovado em todas as disciplinas em que estava matriculado no período letivo; total de situações de trancados - situação que indica que a matrícula está trancada no período letivo.

Outra ação realizada é a criação de três bases com diferentes números de atributos para aplicação dos algoritmos de mineração. Uma por seleção manual de atributos (com 18 atributos) e duas outras escolhidas de acordo com um *ranking* realizado por um algoritmo de seleção de atributos, o *Recursive Feature Elimination* (RFE) [Pedregosa et al. 2011].

A Figura 5 apresenta os dezoito atributos gerados por seleção manual com atributos socieconômicos e acadêmicos dos estudantes e o atributo classe da base de dados (*Sit. da Matrícula*). Já a Figura 6 mostra a classificação gerada por RFE com os dez melhores atributos descritos na Figura 5. Assim, as duas novas bases são geradas com os dez melhores atributos de acordo com a Figura 6 e com os cinco melhores de acordo com o *ranking* apresentado na Figura 6.

Idade	Forma de Ingresso
Sexo	Coeficiente Rendimento Geral
Renda Familiar	Total frequência
Etnia	Total Aprovado
Desc. Estado Civil	Total Aprovado Parcialmente
Desc. Grau Instrução	Total aprovado c/dependência
Estado Civil Pais	Total Reprovado
Grau Instrução Mãe	Total Trancado
Grau Instrução Pai	Sit. da Matrícula
Desc. Turno estudante	

Figura 5. Descrição dos Atributos Selecionados [Fonte: Autores].

1. Total Aprovado	6. Total Reprovado
2. Total aprovado c/dependência	7. Coeficiente Rendimento Geral
3. Total Trancado	8. Desc Turno estudante
4. Total Aprovado Parcialmente	9. Grau Instrução Mãe
5. Total frequência	10. Forma de Ingresso

Figura 6. Descrição dos Dez Melhores Atributos Selecionados por RFE [Fonte: Autores].

5.4. Mineração de Dados

Para a realização da fase de treinamentos e testes com o objetivo de classificar um estudante entre egresso com êxito e egresso sem êxito, são utilizados os algoritmos *Naive Bayes*, KNN (*K-Nearest Neighbor*), Árvore de Decisão, *Random Forest*, Redes Neurais Artificiais (ANNs) e Máquinas de Vetores de Suporte (SVM) disponíveis no *sklearn* [Pedregosa et al. 2011]. Para avaliação dos modelos, os algoritmos são treinados e testados 30 vezes por meio da validação cruzada com 10 *folds* para gerar uma estimativa média de acertos e erros de cada base.

Com exceção do algoritmo *Naive Bayes*, os principais parâmetros dos demais algoritmos são modificados e testados 30 vezes utilizando validação cruzada com 10 *folds* com o objetivo de encontrar os melhores parâmetros para cada algoritmo. A Tabela 1 apresenta os algoritmos com os principais parâmetros utilizados e também os parâmetros

testados com variação. Os demais parâmetros de cada algoritmo são os originalmente disponibilizados no *sklearn*.

Tabela 1. Parâmetros dos Algoritmos [Fonte: Autores]

Algoritmo	**Parâmetros**
KNN	*n_neighbors* = 1 a 50 *metric* = *standard Euclidean metric*
Árvore de Decisão	*criterion* = *entropy*
Random Forest	*n_estimators* =[10,20,30,40,50,100,150,200] *criterion* = *entropy*
ANNs	*activation* = [*logistic sigmoid, hyperbolic tan*] *max_iter* = 1000, *tol* = 0.0001, *solver* = *'adam'*, *learning_rate_init*=0.001, *hidden_layer_sizes*=100
SVM	*kernel* = [*linear, poly, rbf, sigmoid*] *C* = 2.0 (P*enalty parameter C of the error term*)

5.5. Pós-processamento

Na Tabela 2 são apresentados os valores de acurácia média com seu desvio padrão para cada algoritmo em base de dados com diferente quantidade de atributos. A *base1* possui os atributos gerados por seleção manual (Figura 5). A *base2* possui os dez 10 melhores atributos de acordo com o *ranking* gerado pelo algoritmo RFE (Figura 6). A *base3* possui os 5 melhores atributos de acordo com o *ranking* apresentado na Figura 6.

Tabela 2. Resultados dos Modelos Preditivos [Fonte: Autores].

Algoritmo	*base1*	*base2*	*base3*
Naive Bayes	0,9650 0,0021	0,9652 0,0027	0,9731 0,0010
KNN	0,9151 0,0062	0,9643 0,0035	0,9717 0,0044
Árvore de Decisão	0,9665 0,0057	0,9679 0,0049	0,9681 0,0057
Random Forest	0,9665 0,0021	0,9697 0,0031	0,9745 0,0032
ANNs	0,9747 0,0039	0,9762 0,0026	0,9747 0,0019
SVM	**0,9774** **0,0028**	**0,9797** **0,0027**	**0,9750** **0,0023**

Em relação às variações de parâmetros dos algoritmos da Tabela 1, o valor de *n_neighbors* igual a 1 obteve os melhores valores de acurácia para o algoritmo KNN durante os testes para as três bases. Para o algoritmo *Random Forest*, os melhores resultados são com números de *n_estimators* de 150 para a *base1*, 40 para a *base2* e 100 para a *base3*. Já para o algoritmo de ANNs o melhor valor de *activation* é igual a *logistic sigmoid* para

as três bases. Para o SVM, o parâmetro de *kernel* igual a *linear* é o melhor para todas as bases.

De forma geral, o algoritmo *Support Vector Machine* (SVM) obteve os melhores resultados nas três bases testadas, tendo a maior taxa de acerto de 97.97% na *base2* com 10 atributos. Assim, com as características selecionadas em conjunto com o SVM, o algoritmo alcançou uma taxa de acerto melhor que os demais trabalhos relacionados detalhados na Seção 3 em classificar a situação de estudantes. Outro destaque é o algoritmo ANNs, que também na *base2* obteve resultados similares ao SVM. Além disso, pode-se observar que para todos os algoritmos testados a acurácia é superior a 91%, sendo capazes de classificar estudantes egressos com êxito e sem êxito com alta acurácia. Em termos de descoberta de conhecimento em relação aos dados analisados neste trabalho, pode-se ressaltar que os atributos mais importantes na predição de um estudante evadir ou não, são os apresentados na Figura 6 que, em sua maioria, estão relacionados às informações acadêmicas dos estudantes.

6. Conclusões

Mineração de dados educacionais é definida como uma área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Nesse trabalho, são utilizados métodos de mineração de dados e de aprendizagem de máquina em dados acadêmicos e socioeconômicos de estudantes do IFCE para realizar previsão de estudantes com características de evasão ou não. Diante desses resultados, fica claro que o uso de técnicas de mineração de dados em contextos educacionais oferece oportunidades para que educadores e pesquisadores tenham acesso a conhecimentos úteis, gerados a partir de conjuntos de dados de instituições de ensino. Assim, por meio dos resultados das técnicas e algoritmos de mineração de dados, professores e gestores podem descobrir quais fatores estão prejudicando o desempenho de estudantes e, assim, identificar aqueles que estão em risco de evasão ou com baixo desempenho; personalizar e adaptar o conteúdo, as metodologias e os processos avaliativos para atender às necessidades individuais e, melhorar/otimizar o uso dos recursos educacionais.

Como trabalhos futuros, pretende-se utilizar os modelos treinados para classificação de risco de evasão de estudantes já matriculados e de novos ingressantes com objetivo de realizar um acompanhamento dos estudantes ao longo dos semestres. Além disso, objetiva-se ampliar o campo de aplicação da proposta deste trabalho em outros *campi* da instituição, nos diferentes níveis, modalidades e áreas de formação. Pretende-se, ainda, melhorar a base de dados atual com informações dos estudantes em relação às disciplinas cursadas, com o objetivo de melhorar ainda mais a acurácia dos modelos de aprendizagem de máquina.

Referências

BRASIL (1988). *Constituição da República Federativa do Brasil*. Constituição (1988), Senado Federal.

BRASIL (1996). *Lei nº 9.394, de 1996, que estabelece as diretrizes e bases da educação nacional, e legislação correlata*. BRASIL. Lei de Diretrizes e Bases da Educação Nacional.

BRASIL (2014). *Documento orientador para a superação da evasão e retenção na Rede Federal de Educação Profissional, Científica e Tecnológica*. BRASIL. Ministério da Educação.

Correia, E. d. S., da Silva, V. A., and TAVARES, A. C. D. M. (2016). Avaliação da aprendizagem: Do castigo ao diagnóstico pelo professor. *Interfaces Científicas-Educação*, 5(1):21–28.

De Castro, L. N. and Ferrari, D. G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*. Saraiva, São Paulo, 1 edition.

Depresbiteris, L. (1998). Avaliação da aprendizagem do ponto de vista técnico-científico e filosófico-político. *São Paulo: FDE*, pages 161–172.

Devasia, T., Vinushree, T. P., and Hegde, V. (2016). Prediction of students performance using educational data mining. In *International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 91–95.

Dharmawan, T., Ginardi, H., and Munif, A. (2018). Dropout detection using non-academic data. In *4th International Conference on Science and Technology (ICST)*, pages 1–4.

Duda, H. and Hart, P. (2001). *Stork, Pattern Classification*. John Wiley & Sons.

Faceli, K., Lorena, A. C., Gama, J., and de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. LTC, Rio de Janeiro.

Gonçalves, T., Silva, J., and Cortes, O. (2018). Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Computação Aplicada*, 10(3):11–20.

Hegde, V. and Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *2nd International Conference on Inventive Systems and Control (ICISC)*, pages 694–699.

IFCE (2017). *Plano estratégico para permanência e êxito dos estudantes do IFCE*. IFCE. Instituto Federal de Educação, Ciência e Tecnologia do Ceará. Pró-reitoria de Ensino - PROEN, Fortaleza.

IFCE (2018). IFCE em Números. `http://ifceemnumeros.ifce.edu.br/`. [Último Acesso em: 01-Mar-2019].

INEP (2017). *Indicadores de Fluxo Escolar da Educação Básica*. INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília-DF.

Lanes, M. d. A. and Alcântara, C. d. S. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In *XXIX Simpósio Brasileiro de Informática na Educação (SBIE)*.

Lei, C. and Li, K. F. (2015). Academic performance predictors. In *IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, pages 577–581.

Maria, W., Damiani, J. L., and Pereira, M. (2016). Rede bayesiana para previsão de evasão escolar. In *V Congresso Brasileiro de Informática na Educação (CBIE)*, pages 920 – 929.

Marwaha, A. and Ahuja, S. (2017). A review on identifying influencing factors and data mining techniques best suited for analyzing students' performance. In *International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pages 373–378.

Mitchell, T. (1997). *Machine learning*. McGraw Hill, 1 edition.

Paz, F. J. and Cazella, S. C. (2017). Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In *Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (CBIE 2017)*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perez, B., Castellanos, C., and Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. In *IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, pages 1–6.

PNUD (2015). *Relatório do Desenvolvimento Humano de 2015*. PNUD. Programa das Nações Unidas para o Desenvolvimento.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à mineração de dados: com aplicações em R*. Elsevier, Rio de Janeiro, 1 edition.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. In *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6.

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research*, 14:207–226.

Veiga, L. d., Drehmer, C. L., Urnau, J. R., Silva, T. d., Lizote, S. A., and Terres, J. C. (2012). O QUE É UMA UNIVERSIDADE COMUNITÁRIA? um estudo sobre o grau de conhecimento dos estudantes de uma instituição de ensino superior. In *XII Colóquio Internacional sobre Gestão Universitária nas Américas*, pages 1–15.

Witten, I. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

Yukselturk, E., Ozekes, S., and Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning*, 17(1):118–133.

Zhang, L. and Li, K. F. (2018). Education analytics: Challenges and approaches. In *32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 193–198.