Métodos de Machine Learning Aplicados no Cenário da Educação a Distância Brasileira

Charles Nicollas C. Freitas
Departamento de Estatística e Informática - DEINFO
Universidade Federal Rural de Pernambuco - UFRPE
Recife-PE, Brasil
cnicollas21@hotmail.com

Roberta M. M. Gouveia
DEINFO - UFRPE
Recife-PE, Brasil
roberta.gouveia@ufrpe.br

Rodrigo G. F. Soares
DEINFO - UFRPE
Recife-PE, Brasil
rodrigo.gfsoares@ufrpe.br

CEUR-WS.org/Vol-2733/old_ver_proceedings/paper20.pdf

***Abstract*—Tackling student evasion has been a major challenge for the Brazilian Educational System. In the last few years, there has been an increasing interest in Distance Education to address such an important issue. This new paradigm facilitates the attendance of students who have difficulties in attending classes in person due to work, geographical or socioeconomic reasons. However, Distance Education has also shown a growing number of evading students. To tackle Distance Education evasion, we propose the use of Data Mining and Machine Learning techniques to predict the number of students at risk of evasion. Such an approach might help Education Institutions to quantify, plan and develop solutions to this problem. Our work employs Decision Trees, Bootstrap Aggregating Ensemble, Multilayer Perceptron and Support Vector Machines to produce accurate estimates of evading students in Higher Education Institutions that have Distance Education programs. Our experiments showed that our approach could deliver good generalization performance.**

***Resumo*—Combater a evasão de alunos tem sido um grande desafio para o Sistema Educacional Brasileiro. Nos últimos anos, tem havido um interesse crescente pela Educação a Distância para lidar com tal problema. Esse novo paradigma facilita o atendimento a alunos que têm dificuldade em frequentar as aulas presencialmente por motivos laborais, geográficos ou socioeconômicos. No entanto, a Educação a Distância também tem mostrado um número crescente de alunos evadidos. Para enfrentar a evasão na Educação a Distância, propomos o uso de técnicas de Data Mining e Machine Learning para prever o número de alunos em risco de evasão. Tal abordagem pode ajudar as instituições de ensino a quantificar, planejar e desenvolver soluções para este problema. Nosso trabalho emprega Árvores de Decisão, Bootstrap Aggregating Ensemble, Multilayer Perceptron e Support Vector Machines para produzir estimativas precisas de evasão de alunos em instituições de ensino superior com programas de Educação a Distância. Nossos experimentos mostraram que nossa abordagem pode fornecer um bom desempenho de generalização para a predição de evasão.**

***Index Terms*—Educational Data Mining, Machine Learning, Knowledge Discovery in Databases, Educational Systems, Algorithms**

I. Introdução

As tecnologias de informação e comunicação, quando bem utilizadas, tornam-se um diferencial para instituições educacionais que buscam excelência em sua atuação. Este artigo insere-se nas áreas interdisciplinares de *Data Science*, Mineração de Dados Educacionais - MDE, do inglês *Educational Data Mining* - EDM, *Machine Learning* (ML), Banco de Dados, estatística, dentre outras que compõem a base de conhecimento utilizada na análise de dados educacionais.

Este trabalho aplica o processo *Knowledge Discovery in Databases* (KDD), também conhecido como Descoberta de Conhecimento em Bases de Dados, com intuito de encontrar padrões de comportamento e descobrir novos conhecimentos em bases de dados educacionais. A motivação do estudo surge do interesse em adquirir regras significativas, na tentativa de melhor compreender algumas adversidades da educação superior, enfrentados na modalidade a distância. Assim, os resultados desse estudo podem ser úteis para profissionais envolvidos com a implementação de métodos de Mineração de Dados - MD, do inglês *Data Mining*, no contexto da Educação a Distância (EaD).

Inicialmente foi realizado um levantamento acerca de trabalhos relacionados com EaD e as áreas interdisciplinares elencadas acima. Em seguida, focou-se na obtenção e tratamento dos dados, dando seguimento à etapa de pré-processamento, finalizando com aplicação de técnicas e algoritmos de *Data Mining* para descoberta de novos conhecimentos e detecção de padrões nos dados. Diante desse contexto, o objetivo do trabalho consiste em aplicar o processo KDD para traçar o perfil da EaD em uma universidade pública brasileira, com vista à obtenção de um melhor entendimento acerca de estudantes e cursos realizados em ambientes *e-learning*.

A busca por uma educação além do limite espaço-tempo, que visa transformar e evoluir o processo tradicional de aprendizagem, é uma das propostas da EaD [1]. O conceito formal de EaD, definido pelo Secretaria de Educação Superior (SESu) do Ministério da Educação (MEC), está presente no Decreto no 5.622, 19.12.2005, que regulamenta o Art. 80 da Lei 9394/96, Lei de Diretrizes e Bases da Educação Nacional - LDB. De acordo com o MEC, a EaD é definida como [2]: "A modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempos diversos."

A internet e os *softwares* educacionais de suporte ao processo de ensino-aprendizagem surgiram como potencializadores da EaD, dando início ao termo *e-learning*, ou

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

aprendizagem eletrônica, que especifica a EaD realizada por meio de plataformas computacionais e Ambientes Virtuais de Aprendizagem - AVA. Vale destacar que existe uma sutil diferença entre os termos EaD e *e-learning*, já que a EaD pode ser realizada sem o suporte eletrônico, enquanto *e-learning* necessita do suporte eletrônico [3]. Assim, nesse trabalho é utilizado o termo Educação a Distância de forma genérica, tanto para referenciar a EaD tradicional (sem o suporte eletrônico), como *e-learning* (com suporte eletrônico), ou seja, EaD designando a modalidade de ensino a distância independentemente da mídia que a suporta.

Embora tenha aumentado o número de instituições educacionais que aderiram à EaD em seus cursos de graduação e especializações, o Brasil ainda está em fase de transição nessa modalidade, visto que algumas delas estão se limitando a reproduzir para o ambiente virtual pequenas adaptações do ensino presencial. Em alguns casos, as aulas são disponibilizadas do ensino presencial para o virtual sem qualquer alteração didático-pedagógica nos processos de ensino-aprendizagem. Essas práticas contribuem para aumentar os índices de evasão e retenção dos estudantes.

A Mineração de Dados Educacionais utiliza técnicas de MD para explorar dados oriundos de contextos educacionais, sendo aplicada nos seguintes domínios: (I) Educação *Offline*: para análises de dados de desempenho e comportamento dos estudantes, bem como análises de currículo/histórico escolar, ou seja, dados gerados em ambientes de sala de aula; (II) Aprendizagem Eletrônica, mais conhecida como *e-learning*, e Sistema de Gestão da Aprendizagem, do inglês *Learning Management System* - LMS: para análise de dados armazenados em sistemas LMS no formato de *logs* e bases de dados; (III) Sistemas Tutores Inteligentes, do inglês *Intelligent Tutoring System* - ITS, e Sistemas Hipermídias Adaptativos Educacionais, do inglês *Adaptive Educational Hypermedia System*: os quais são aplicados sobre dados de sistemas que se adaptam ao percurso de cada estudante no ambiente virtual de aprendizagem [4].

Conforme ilustra a "Fig. 1", a EDM é a combinação de 3 (três) principais áreas de conhecimento: Ciência da Computação, Educação e Estatística. A interseção dessas áreas fornece três subáreas, que são: *e-learning*, *Data Mining* e *Machine Learning*, e *Learning Analytics*.

A área interdisciplinar de Mineração de Dados Educacionais vem se consolidando na última década, tendo vários *papers* publicados em revistas e conferências relevantes. Alguns pesquisadores realizaram levantamentos detalhados acerca da MDE, sendo fontes de referências recomendadas [4], [5], [6], [7], [8]. Na literatura existem vários trabalhos relacionados à aplicação de técnicas de MD e ML no contexto educacional [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. Esses artigos trazem excelentes contribuições sobre aplicações de algoritmos de mineração de dados, tanto no cenário da educação presencial, quanto em ambientes virtuais de aprendizagem da educação a distância. São reflexões fundamentadas sobre os desafios da educação, especialmente em instituições públicas de ensino superior.

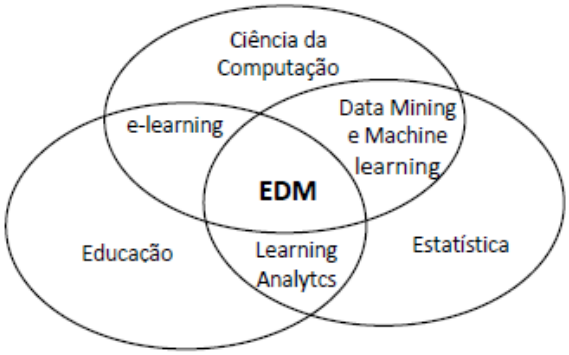

Fig. 1. Principais áreas relacionadas com EDM.

O artigo está organizado como segue: a seção 2 tem o objetivo de contextualizar o problema e os métodos de *Machine Learning* usados no trabalho. Na seção 3 são descritos os experimentos realizados. A seção 4 apresenta os resultados e suas respectivas análises. Por fim, as conclusões e possíveis trabalhos futuros são apresentados na seção 5.

II. Metodologia

A aplicação do *Data Mining* visa encontrar o perfil do estudante e detectar ineficiências da EaD, que por sua vez desestimulam os alunos a prosseguirem nos cursos. Os resultados obtidos com *Data Mining* são utilizados a fim de detectar padrões, descobrir regras significativas e estabelecer relações entre os índices de evasão e retenção, o perfil socioeconômico dos alunos e as características inerentes da EaD. Ao constatar tais relacionamentos e pontos fracos, ações poderão ser tomadas, por parte da instituição, para eliminá-las, buscando reduzir os altos índices de evasão constatados na modalidade a distância.

A principal relevância da pesquisa no desenvolvimento científico e tecnológico refere-se ao fato do trabalho propor uma análise do cenário da EaD, por meio do processo computacional de descoberta de conhecimento em bases de dados, utilizando técnicas de classificação de padrões.

Foram obtidos dados acadêmicos de uma Instituição Federal de Ensino Superior (IFES) para análise pontual e concreta da educação a distância. Os dados dos estudantes da EaD referem-se a um período de 8 anos, e foram obtidos nos formatos *txt* e *xlsx*, sendo em seguida, consolidados em um arquivo *csv*. A pesquisa é baseada no anonimato, por isso não foram obtidas informações como nome e CPF, visando preservar as identidades dos alunos. Foram obtidos dados dos seguintes cursos: Licenciatura em Letras; Licenciatura em Pedagogia; Licenciatura em Computação e Bacharelado em Administração Pública.

Os dados obtidos foram de contexto histórico, para uma análise distintiva da evolução da EaD na instituição, sendo obtidos as seguintes informações: (I) Histórico Escolar; (II) Dados gerais sobre o aluno, tais como: Forma de ingresso; Período de ingresso; Curso; Área de Conhecimento; Polo; Modalidade (licenciatura, bacharelado, tecnólogo etc.); Idade;

Gênero/Sexo; Estado Civil; Naturalidade; Nacionalidade; Etnia/Raça (cor da pele); Deficiência; Situação Acadêmica (Cursando, Concluído, Abandono etc.); (III) Dados Socioeconômicos, tais como informações sobre ensino fundamental e médio (tipo de escola); Renda familiar; Trabalho remunerado; Se possui computador em casa; Acesso à internet etc.

Esses dados são essenciais para compreender quais são as potenciais deficiências e obstáculos enfrentados pelos docentes, estudantes e gestores, como também ter um entendimento sobre os estudantes da modalidade a distância em uma IFES, afinal o principal objetivo é obter um respaldo científico necessário para detectar padrões e descobrir regras significativas sobre os índices de evasão e retenção em cursos a distância, adquirindo um melhor entendimento acerca da EaD no cenário brasileiro.

A. Métodos de Machine Learning

Os algoritmos de *Data Mining* interpretam os dados a fim de produzir uma quantidade de padrões úteis, válidos e de fácil entendimento. Os resultados gerados podem ser usados para predições e têm por finalidade conduzir a tomadas de decisões inteligentes. O fator humano faz parte de todo o processo, por isso não pode ser uma ação totalmente automatizada.

Os algoritmos de mineração de dados favorecem a extração de informações de grandes volumes de dados, e a análise estatística desses dados permite que se observem tendências e respostas para situações diversos, tais como: encontrar e detectar cursos onde as evasões são mais frequentes; determinar perfis (comportamentos típicos), e associar categorias de alunos e cursos com características de sucesso na EaD; elencar dificuldades frequentemente enfrentadas pelos docentes e discentes da EaD; identificar nos AVAs as disciplinas com alto índice de reprovação e suas causas etc.

Alguns pré-requisitos são essenciais para o sucesso da mineração de dados, por isso foram construídos modelos baseados em metas preditivas e descritivas. Diante das metas preditivas, tem-se, por exemplo, a utilização da tarefa de Classificação por Árvore de Decisão.

Dentre os vários métodos de *Machine Learning* disponíveis na literatura, cinco deles se mostram adequados aos resultados pretendidos por este estudo. Os métodos aplicados foram: Classificação por Árvore de Decisão, Classificação Bayesiana, Classificação por Redes Neurais, Classificação por *Ensembles* e Classificação por *Support Vector Machine* - SVM, todos inerentes ao Aprendizado Supervisionado. Os algoritmos de classificação utilizados foram: NaiveBayes, J48 (árvore de decisão), MultilayerPerceptron - que implementa o backpropagation para classificação (Redes Neurais), LibSVM (SVM), Bagging e AdaBoost (Ensembles).

A Classificação Bayesiana (*Bayesian Classification*) é uma técnica estatística (probabilidade condicional) baseada no teorema de Thomas Bayes. Segundo o teorema de Bayes, é possível encontrar a probabilidade de certo evento ocorrer, dada a probabilidade de outro evento que já ocorreu. Comparativos mostram que os algoritmos Bayesianos, chamados de *Naive Bayes*, obtiveram resultados compatíveis com os métodos de árvore de decisão e redes neurais. Devido a sua simplicidade e o alto poder preditivo, é um dos algoritmos mais utilizados. O algoritmo *Naive Bayes* parte do princípio que não exista relação de dependência entre os atributos, no entanto, nem sempre isto é possível [22].

A técnica de Redes Neurais é muito utilizada em tarefas de classificação, regressão e segmentação. Os dados são trabalhados com base no funcionamento do cérebro humano, aprendendo a tomar decisões baseadas nas experiências anteriores (nas instâncias anteriores dos dados). Os neurônios do cérebro são representados por nodos que estão conectados em outros nodos por sinapses, formando uma rede de processamento. Os valores das entradas são multiplicados nos neurônios pelos pesos de suas sinapses, conforme vão percorrendo a rede. Ao final, temos uma classificação ou a previsão da entrada [23].

As árvores de decisão têm como objetivo principal dividir as instâncias em classes. Cada nó da árvore testa o domínio de uma variável de entrada e o redireciona para o nó seguinte. Cada sub-árvore representa o resultado de um teste e a folha é a classificação que aquele registro recebeu. Ao final, cada nó terminal terá os registros da entrada que se adéquam às regras regidas por esse nó, representando assim, uma classe [23].

Os classificadores ensembles, comitê de especialistas, predizem a classe de um registro elegendo a maioria dos votos feitos pelos classificadores base. Para isso, deve-se evitar: subconjuntos idênticos (os erros serão os mesmos), e subconjuntos disjuntos (erros não correlacionados). Para que a performance de um método ensemble seja melhor que a de um classificador simples, os classificadores base devem ser independentes, e devem ter performance melhor que um *random guessing* [11]. Os dois tipos de ensembles escolhidos neste trabalho foram: *Bagging* e *Adaboost*.

O *Support Vector Machines* é baseado no conceito de planos de decisão que definem limites de decisão (Vetor Suporte). Um plano de decisão separa um conjunto de objetos com diferentes associações de classe. SVM é essencialmente um método classificação que executa tarefas de classificação através da construção de hiperplanos em um espaço multidimensional que separa casos de diferentes rótulos de classe. Ele suporta ambas as tarefas de regressão e de classificação e pode lidar com múltiplas variáveis contínuas. Para construir um hiperplano ótimo, o SVM emprega um algoritmo iterativo de formação, que é usado para minimizar uma função de erro [24].

Há um número de núcleos que podem ser usados em modelos *Support Vector Machines*. Estes incluem linear, polinomial, função *radial base* (RBF) e *sigmóide*. Estas funções de núcleo representam um produto de ponto de pontos de dados de entrada mapeado para o maior espaço de características dimensionais por transformação.

Nem todas as regras geradas pelo *Data Mining* são consideradas relevantes para o processo de extração do conhecimento em banco de dados, visto que o especialista precisa interpretá-las no contexto em que está inserido e só depois aplicá-las, afinal o fator humano também faz parte do processo. Desta forma, o especialista do negócio precisa avaliar as regras para

que o resultado seja aplicável na prática.

III. Experimentos

Com o objetivo de verificar a adequação do conjunto de dados propostos, foram realizados experimentos com a base de dados citada anteriormente, contendo informações de estudantes em quatro cursos realizado a distância. Foram desenvolvidos procedimentos para extração dos atributos considerados significativos para este trabalho.

A. Pré-Processamento

Para se ter uma visão geral preliminar dos dados, se configura uma boa prática fazer inicialmente uma análise descritiva dos dados, também conhecida como análise exploratória dos dados. Neste diagnóstico inicial, medições são feitas sobre os atributos dos dados como média/mediana, desvio padrão, valor mínimo, máximo, *outliers*, entre outros . Estas medidas auxiliam no encaminhamento da solução de pré-processamento a ser adotada e também, em caso de valores ausentes já será possível verificar a sua existência e, consequentemente a sua solução.

Valores ausentes, ou *missing values* , são atributos que não tem valores preenchidos. O tratamento pode ser feito pela simples remoção do atributo (em caso de grande incidência) ou do exemplar (em caso de poucas ocorrências). Ou ainda o valor pode ser substituído por uma constante calculara pela média, mediana, valor máximo ou mínimo. Outro tratamento que pode ser diagnosticado na análise descritiva são os valores ruidosos ou que estão fora do padrão (*outliers*). Este tipo de situação ocorre quando surge algum exemplar com valor de atributo que foge de um padrão. Por fim, um cenário que surge tipicamente quando se faz integração de dados é a inconsistência de valores. A inconsistência ocorre quando há falta de um critério bem definido entre os valores dos atributos ou dos exemplares.

A normalização de valores consiste em uma técnica para deixar os valores dos atributos em uma mesma escala. Abordagem de solução comum é calcular o valor máximo de um atributo para dividi-lo aos demais exemplares com mesmo atributo. A normalização faz parte de um tratamento chamado transformação de valores que ainda compreende a mudança de tipos categóricos para numéricos. No caso do gênero, por exemplo, atributo nominal, como são apenas dois valores, eles poderiam ser transformados para binário 0 e 1. No entanto, deve-se ter cuidado para não transformar um atributo nominal em ordinal no processo de transformação, isto é, o valor não pode ideia de ordem.

Finalmente, foi feita a seleção de atributos na fase de pré-processamento. Dentre as causas que levam a se fazer este tipo de análise, tem-se: integração de bases, falta de definição clara de atributos que representam um problema, grande disponibilidade de dados e outras. A seleção consiste basicamente em escolher o melhor conjunto de dados que representam a base original com a mesma capacidade analítica.

Para a realização da seleção de exemplares foi utilizado o método *Classifier Subset Evaluator* (CSE). Este método permite avaliar subconjuntos de atributos em dados de treinamento ou um conjunto de testes independente. Utiliza um classificador para estimar a ”mérito” de um conjunto de atributos. Junto ao CSE foi utilizado um método de pesquisa de atributos *BestFirst*, que auxilia na busca por um subconjunto de atributos que represente a base original. Foi escolhida a direção *Forward*, que começa com o conjunto vazio de atributos e procura para frente, considerando todas as possíveis adições de atributos individuais e deleções em um determinado ponto, no caso o *searchTermination* que é o parâmetro de parada do método, sendo o valor 5 escolhido nesse trabalho.

Após a realização da etapa anterior, foram selecionados 20 atributos dos 214 da base original, isto significa que estes atributos representam melhor a base original em termos de generalização do problema. Dentre os atributos selecionados, tem-se as seguintes informações sobre os estudante: nome do curso, área de conhecimento, polo, ano de ingresso, status acadêmico, estado civil, idade, naturalidade (estado), tipo de deficiência, tipo de escola do ensino médio, tipo de escola do ensino fundamental, se possui internet, se possui trabalho remunerado e média geral.

B. Avaliação dos Modelos de Machine Learning

O objetivo principal do experimento é verificar a relevância dos atributos elencados acima, bem como analisar o impacto da aplicação de técnica de seleção de atributos na acurácia da previsão de desempenho dos seis classificadores. A acurácia é a proporção entre o número de estudantes corretamente classificados pelos algoritmos em sua respectiva classe, e o número total de estudantes considerados no estudo.

Para o desenvolvimento deste trabalho foram utilizados seis algoritmos de classificação, como descritos anteriormente, que são eles: *MultilayerPerceptron* (MLP), *NaiveBayes*, SVM, J48, *Adaboost* e *Bagging*. Para auxiliar na avaliação dos resultados e o cálculo da acurácia utilizou-se o método *K-fold Cross-Validation*, que consiste em uma técnica para a estratificação da base dados em conjunto de treinamento e teste. Geralmente, sugere-se a adoção de k igual a 10 como valor padrão para o número de partições dos dados [25].

O primeiro experimento corresponde à seleção dos melhores parâmetros definidos a priori para cada um dos seis algoritmos. O processo de avaliação de desempenho de cada combinação de parâmetros se baseia no método descrito anteriormente, sendo 30 parâmetros para *MLP*, *SVM* e *Bagging*, 18 parâmetros para J48, 6 parâmetros para *Adaboost*, e nenhum parâmetro para *NaiveBayes*. Assim é possível definir quais são os melhores parâmetros de cada algoritmo, para posteriormente definir qual a porcentagem de assertividade de cada um deles.

Os gráficos apresentados na “Fig. 2” destacam os resultados obtidos no experimento de seleção de parâmetros por modelo, demonstrando a precisão dos algoritmos para predizer o desempenho em cada combinação de parâmetros. O resultado apresentado, no eixo vertical, corresponde a um valor médio obtido para as combinações de parâmetros, onde foram se-

lecionados os melhores parâmetros de cada algoritmo para a realização do segundo experimento.

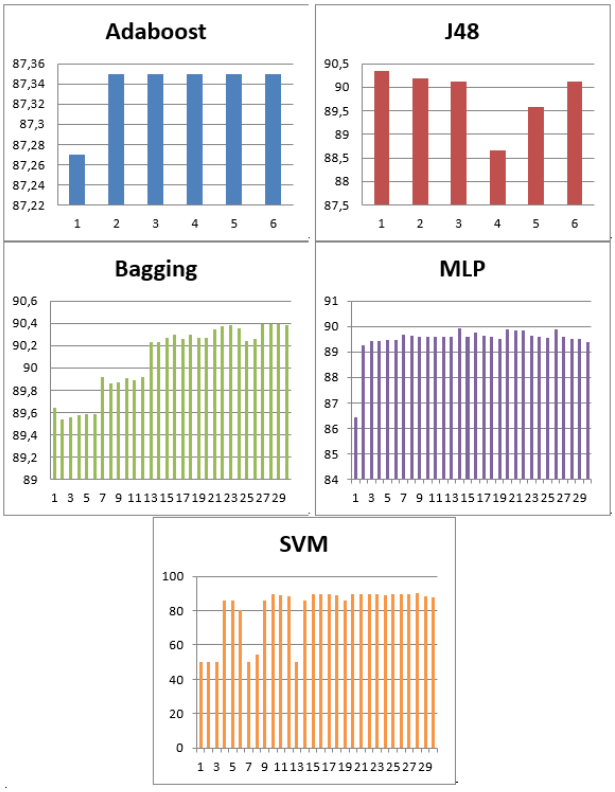

Fig. 2. Experimento de seleção de parâmetros por modelo.

Para o experimento 2, foram utilizados todos os melhores parâmetros obtidos no experimento 1. Com o objetivo de testar a significância estatística dos resultados obtidos, utilizou-se a técnica de teste estatístico *pair-wise T-Test* [25], com nível significância de 5%.

A Tabela I apresenta o resumo com os resultados do segundo experimento executado. Nela, constam o percentual de acurácia médio e o desvio padrão dos seis algoritmos analisados. Como o desvio padrão refere-se a quantidade de variação (dispersão) dos dados dentro da amostra em relação à média, então um baixo valor para desvio padrão indica que a amostra tende a ser mais homogênea.

TABLE I
ACURÁCIA MÉDIA E DESVIO PADRÃO DOS CLASSIFICADORES.

	Adaboost	J48	Bagging
Acurácia (Desvio Padrão)	87,35 (0,59)	90,35 (0,28)	90,33 (0,31)

	MLP	SVM	NaiveBayes
Acurácia (Desvio Padrão)	89,85 (0,57)	89,91 (0,38)	89,31 (0,56)

IV. Análise de Resultados

Avaliando os resultados obtidos percebe-se que a utilização do conjunto completo de atributos proposto, juntamente com a técnica de otimização de parâmetros empregada no experimento 1, obteve os melhores resultados em termos da taxa acurácia. Destaca-se que nos seis classificadores utilizados neste experimento esta tendência pode ser observada.

Um aspecto a ser destacado, a partir dos testes realizados, aponta para a viabilidade da utilização de um conjunto amplo de atributos para representação do perfil dos estudantes, potencialmente generalizáveis a diversos cenários de cursos EAD.

Tomando-se como base o experimento 2, observou-se que o algoritmo J48 apresentou melhor classificação, com taxa de acerto de 90,35% e 0,28 de desvio padrão. Já o algoritmo *Adaboost* apresentou menor acurácia, 87,35%, e maior desvio padrão (0,59). Os resultados obtidos demonstram que os algoritmos *Bagging*, *MultilayerPercepetron*, J48 e SVM podem ser utilizados para realizar inferências em relação aos índices de evasão dos alunos, por possuírem taxa de acurácia acima da média geral (89,51%) de todos os algoritmos analisados.

Apesar do desbalanceamento do atributo classe (Situação Acadêmica), a medida de desempenho utilizada neste trabalho (Taxa de Acurácia) está coerente com as demais métricas da matriz de confusão, a saber: *Precision*, *Recall*, *F-Measure*, e *AUC - Area Under the ROC Curve*.

A partir dos resultados dos experimentos foi possível adquirir o respaldo científico necessário para detectar padrões e descobrir regras significativas na tentativa de melhor compreender a EaD, esta que, por sua vez, exige inovação e infraestrutura tecnológica, além de apoio ao estudante em níveis mais elevados, em comparação à modalidade presencial. Conforme relatório analítico do Censo da EaD no Brasil, 53% dos estudantes brasileiros da modalidade a distância são mulheres, com 39,3% entre 26-30 anos. Aproximadamente 70% das instituições privadas e públicas federais contam com estudantes que, em sua maioria, estudam e trabalham. Em se tratando das taxas de evasão reportadas nos cursos a distância, O Censo da EaD registra uma evasão de 26% a 50% – alertando que a desistência dos estudantes da EaD é maior em comparação aos cursos presenciais. As instituições apontam o fator *tempo* como o mais influente no fenômeno da evasão, seguido do fator *financeiro* [26].

V. Conclusões

O trabalho pretende provocar interesse em instituições, pesquisadores e profissionais envolvidos com a implementação e utilização de sistemas de informações gerenciais de apoio à decisão no contexto da EaD. Tais tecnologias se propõem em fornecer indicadores de qualidade às IFES, proporcionando tomadas de decisões que visam, dentre outras ações, a redução da evasão e retenção de estudantes e, consequentemente, a melhoria da EaD.

A metodologia foi fundamentada no processo KDD, que por sua vez propõe encontrar e interpretar padrões/regras mediante integração de diversas fontes de dados, sendo proposto para determinar as etapas que produzem conhecimentos a partir dos dados e, principalmente, definir a etapa de *Data Mining* [27]. O objetivo é extrair de bases de dados, sem nenhuma

formulação prévia de hipóteses, informações desconhecidas a priori, factíveis, válidas e acionáveis, que poderão ser úteis para a tomada de decisão [28], [29].

Por meio da análise do histórico acadêmico e perfil socioeconômico de estudantes, uma instituição educacional pode ser capaz de acompanhar o rendimento acadêmico do discente, verificando se ele possui potencial para se evadir ou não do curso. Tendo esse conhecimento prévio, as instituições de ensino superior poderão avaliar as necessidades individuais do aluno, e assim, agir de maneira proativa e mais efetiva para que o estudante possa continuar sua graduação.

Com os resultados dos experimentos realizados neste trabalho, é possível a obtenção de indicadores a serem implementados em ambientes virtuais de aprendizagem para a previsão de índice de evasão de estudantes. Estes indicadores podem ser melhorados à medida que a base de dados de treinamento for aumentando. Portanto, o trabalho viabilizou a avaliação de desempenho de seis modelos de *Machine Learning – NaiveBayes*, J48, *MultilayerPerceptron*, *LibSVM*, *Bagging* e *AdaBoost*, com vistas à descoberta de conhecimento no contexto da educação superior brasileira da modalidade a distância.

AGRADECIMENTO

Os autores agradecem o apoio da Fundação de Amparo a Ciência e Tecnologia de Pernambuco - FACEPE, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, e Universidade Federal Rural de Pernambuco - UFRPE.

REFERENCES

[1] J. M. Moran. Educação a distância no brasil: situação e perspectivas, 2014. Disponível em: http://www2.eca.usp.br/moran. Acesso: 18 jun. 2020.

[2] BRASIL. Leis de diretrizes e bases da educação nacional. Decreto n. 5.622, de 19 de dezembro de 2005. Regulamenta o art. 80 da Lei 9.394/96, 20 dez. 1996. Disponível em: http://encurtador.com.br/ckKSZ. Acesso em 08 jun. 2014.

[3] R. M. M. Gouveia. Análises e perspectivas da educação a distância no ensino superior brasileiro. Revista Acesso Livre, p. 207-228, 2017.

[4] C. Romero and S. Ventura. Educational Data Mining: A Review of the State of the Art" in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601-618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[5] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 10, n. 3, p. e1355, 2020, doi: https://doi.org/10.1002/widm.1355

[6] A. Peña-Ayala. Educational data mining: a survey and a data mining-based analysis of recent works. Expert systems with applications, v. 41, p. 1432-1462, 2014, doi: https://doi.org/10.1016/j.eswa.2013.08.042

[7] S. K. Mohamad, Z. Tasir. Educational data mining a review. Procedia Social and Behavioral Sciences, v. 97, 2013, doi: https://doi.org/10.1016/j.sbspro.2013.10.240

[8] H. Aldowah, H. Al-Samarraie, W. M. Fauzy. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37, 13-49, 2019, doi: https://doi.org/10.1016/j.tele.2019.01.007

[9] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, G. V. Erven. Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, v. 94, p. 335-343, 2019, doi: https://doi.org/10.1016/j.jbusres.2018.02.012 https://www.overleaf.com/project/5f0c75b9b4fb520001add8f4

[10] Brandão, J. O. S.; Silva, A. J.; Gouveia, R. M. M.; Soares, R. G. F. Aprendizagem de Máquina para Predição de Desempenho de Estudantes de Graduação na UFPE. In: Brazilian Conference on Intelligent Systems (BRACIS) – XIV Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2017.

[11] E. A. Amrieh, T. Hamtini, I. Aljarah. Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, v. 9, n. 8, p. 119-136, 2016, doi: http://dx.doi.org/10.14257/ijdta.2016.9.8.13

[12] C. N. Freitas, R. M. M. Gouveia, A. Silva. Online Analytical Processing em ambientes virtuais de aprendizagem da educação a distância. In: DesafIE - Workshop de Desafios da Computação Aplicada à Educação – XXXV Congresso da Sociedade Brasileira de Computação, 2015.

[13] L. A. Silva; A. H. Morin; T. M. C. Sato. Práticas de Mineração de Dados no Exame Nacional do Ensino Médio. In: Congresso Brasileiro de Informática na Educação – Workshop de Mineração de Dados em Ambientes Virtuais do Ensino/Aprendizagem, 2014. p. 651-660.

[14] R. Baker, S. Isotani, A. Carvalho. Mineração de dados educacionais: oportunidades para o Brasil. Revista Brasileira de Informática na Educação, v. 19, n. 02, p. 03, 2011, doi:http://dx.doi.org/10.5753/rbie.2011.19.02.03

[15] A. J. C. Kampff. Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. Tese (doutorado), Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Informática na Educação. Porto Alegre/RS/Brasil, 2009.

[16] L. C. Santana; A. M. Maciel; R. L. Rodrigues. Avaliação do perfil de uso no ambiente moodle utilizando técnicas de mineração de dados. In: Simpósio Brasileiro de Informática na Educação, 2014. Congresso Brasileiro de Informática na Educação, 2014.

[17] H. Guércio, P. Marques, V. Ströele, C. K. Pereira, E. Barrere. Análise do desempenho estudantil na educação a distância aplicando técnicas de mineração de dados. In: Congresso Brasileiro de Informática na Educação – Workshop de Mineração de Dados em Ambientes Virtuais de Ensino/Aprendizagem, p. 641-650, 2014

[18] E. Gottardo, C. A. A. Kaestner, R. V. Noronha. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. Revista Brasileira de Informática na Educação, v. 22, n. 01, p. 45, 2014, doi:http://dx.doi.org/10.5753/rbie.2014.22.01.45

[19] S. Singh, V. Kumar. Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education. International Journal of Computer Science and Network - IJCSN, Vol. 1(4), 2012.

[20] M. L. B. Lorenzo, E. G. Sánchez. Predicción de pérdida de implicación de los participantes de un curso en línea masivo y abierto. In: XVIII Simposio Internacional de Informática Educativa - SIIE, 2016.

[21] F. Tanaka, G. Silva, S. Peres, M. Fantinato. Predição de desempenho de alunos no ensino a distância via mineração de processos. In: Brazilian Conference on Intelligent Systems (BRACIS) - XIV Encontro Nacional de Inteligência Artificial e Computacional – ENIAC, 2017.

[22] A. Q. Ayinde, A. B. Adetunji, M. Bello, O. A. Odeniyi. Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data. International Journal of Computer Science Issues - IJCSI, v. 10, n. 4, p. 147, 2013.

[23] T. Devasia, T. P. Vinushree, V. Hegde. Prediction of students performance using Educational Data Mining". International Conference on Data Mining and Advanced Computing - Sapience, IEEE, 2016, doi: 10.1109/SAPIENCE.2016.7684167

[24] D. Ifenthaler, C. Widanapathirana. Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines. Springer - Tech Know Learn 19, 221–240, 2014, doi: https://doi.org/10.1007/s10758-014-9226-4

[25] I. H. Witten, E. Frank, M. A. Hall. Data mining: practical machine learning tools and techniques. 4rd ed. Morgan Kaufmann - Elsevier, 2016.

[26] ABED – Associação Brasileira de Educação a Distância. Censo EAD BR: relatório analítico da aprendizagem a distância no brasil. InterSaberes, 2018.

[27] P. Tan, M. Steinbach, A. Karpatne, V. Kumar. Introduction to Data Mining". 2nd ed. Pearson, 2018.

[28] W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus. Knowledge discovery in databases: An overview. AI magazine, v. 13, n. 3, p. 57, 1992.

[29] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, v. 17, n. 3, p. 37-54, 1996.