2018 | Volume 15 | Nº 2 | Pág. 425 a 438

DOI: http://dx.doi.org/10.15536/thema.15.2018.425-438.868

CIÊNCIAS EXATAS E DA TERRA

Modelo de predição da evasão de estudantes em cursos técnicos a distância a partir da contagem de interações

Prediction model of student dropout in technical distance learning courses considering the interactions count

Emanuel M. Queiroga[1]; João L. Lopes[2]; Ricardo M. de Araujo[3]; Cristian Cechinel[4]

RESUMO

A evasão é considerada um dos principais problemas relacionados com a Educação a Distância (EaD). Na EaD a interação entre estudantes e professores costuma ser mediada por um Ambiente Virtual de Aprendizagem (AVA), sendo gerado um grande volume de dados que fica registrado em arquivos. Assim, este artigo utiliza a mineração de dados para analisar os registros de interações de estudantes dos cursos técnicos a distância, do Câmpus Visconde da Graça (CaVG) do Instituto Federal Sul-rio-grandense (IFSul), com o objetivo de identificar estudantes em risco de evasão. Entende-se como contribuição central deste trabalho a geração de modelos de predição que permitem apoiar as ações de combate à evasão, utilizando como variáveis principais de entrada para os modelos a contagem de interações dos estudantes através do AVA e atributos das mesmas. A premissa é que essa abordagem permite uma maior generalização, possibilitando o emprego em diferentes instituições, cursos e plataformas. Os resultados apontam a possibilidade de predição de estudantes em risco de evasão já nas primeiras semanas dos cursos, com os modelos apresentando taxas de desempenho superiores a 75%.

***Palavras-chave**: educação a distância, evasão de estudantes, modelos de predição, contagem de interações.*

ABSTRACT

Dropout is considered one of the main problems in Distance Learning (DL). In the DL the interaction between students and teachers is usually mediated by a Virtual Learning Environment (VLE), generating a large volume of data that is recorded in archives. Thus, this article uses data mining to analyze the interactions logs of students of the technical distance learning courses, from Campus Visconde da Graça (CaVG) of the Federal Institute Sul-rio-grandense (IFSul), with the objective of identifying students at risk of dropout. The central contribution of this work is the generation of prediction models that support actions to avoid dropout, using as input variables for the models the counting of students' interactions through VLE and their attributes. The premise is that this approach allows a greater generalization, making possible the employment in different institutions, courses and platforms. The results point to the possibility of predicting students at risk of dropout already in the first weeks of the courses, with the models presenting performance rates above 75%.

***Keywords:** distance learning, student dropout, prediction models, interaction counts.*

[1;2] Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense, Pelotas/RS – Brasil.

[3] Universidade Federal de Pelotas, Pelotas/RS – Brasil.

[4] Universidade Federal de Santa Catariana, Araranguá/SC – Brasil.

1. INTRODUÇÃO

A Educação a Distância (EaD) pode ser considerada uma alternativa para o acesso aos programas de qualificação profissional no Brasil (DELANO; CORREA, 2013). Porém, um dos principais desafios da EaD é obter a diminuição da evasão dos alunos, cujo percentual se encontra na faixa de 11% a 25% nos cursos regulamentados pelo Ministério da Educação (MEC), conforme o Censo EAD.BR 2016 (ABED, 2017). Segundo Manhães et al. (2011), a detecção precoce de estudantes com risco de evasão é uma condição importante para redução deste problema, pois possibilita a realização de atendimentos direcionados a situações específicas dos alunos.

O processo de identificação de alunos com risco de evasão, usualmente, é manual, subjetivo, empírico e sujeito a falhas, dependendo especialmente da experiência acadêmica e do envolvimento dos docentes (MANHÃES et al., 2011). Considerando que os docentes desempenham inúmeras atividades, bem como os cursos de EaD normalmente possuem um grande número de alunos, torna-se bastante difícil acompanhar e reconhecer as necessidades de cada estudante e identificar aqueles que apresentam risco de evasão.

Nesse cenário, a mineração de dados mostra-se bastante oportuna para o tratamento e a descoberta de conhecimento no grande volume de dados gerado pelas interações dos estudantes nos Ambientes Virtuais de Aprendizagem (AVAs). Um AVA pode ser considerado o "local virtual" onde os cursos na modalidade a distância normalmente ocorrem. De modo geral, os AVAs oferecem áreas para apresentação de conteúdos em diversos formatos, além de atividades de verificação da aprendizagem e espaços para interação síncrona e assíncrona. Um AVA também mantém registro em seus arquivos de todas as interações virtuais que ocorrem ao longo do processo de ensino-aprendizagem.

Nesse sentido, a Mineração de Dados Educacionais (EDM - Educational Data Mining) vem mostrando um grande potencial para possibilitar a melhora da qualidade do ensino (BAKER et al., 2011). Essa área de pesquisa pode prover informações para as instituições de ensino criarem modelos de predição que permitam avaliar o risco de evasão dos estudantes. A aplicação da EDM pode possibilitar o tratamento diferenciado dos alunos pelo emprego de estratégias de apoio específicas aos alunos que estejam com maior probabilidade de evasão.

Considerando o exposto, o presente trabalho define uma abordagem para a detecção de estudantes em risco de evasão em cursos técnicos a distância do Câmpus Visconde da Graça (CaVG) do Instituto Federal Sul-rio-grandense (IFSul). Para tanto, a metodologia empregada considera apenas a contagem de interações dos estudantes no AVA e atributos derivados dessas contagens. A premissa é que essa abordagem permite uma maior generalização, possibilitando o emprego em diferentes instituições, cursos e plataformas, uma vez que não utiliza diferenciações entre os diversos tipos de interações, quão menos faz uso de informações adicionais que estejam fora do ambiente virtual, como dado geográficos ou econômicos.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta a fundamentação teórica do trabalho. A Seção 3 apresenta uma sistematização de trabalhos relacionados à predição de estudantes em risco de evasão. Na Seção 4 são descritos os dados e a metodologia utilizada nos experimentos realizados. A Seção 5 discute os resultados obtidos. Por fim, a Seção 6 apresenta as conclusões do trabalho e as possibilidades de trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

No século XXI as instituições de ensino devem estar preparadas para interagir com uma geração mais atualizada e informada. Os atuais meios de comunicação, especialmente a Internet, permitem o acesso instantâneo à informação e os alunos têm mais facilidade para buscar conhecimento por meio da tecnologia colocada à sua disposição (FARIA, 2004).

Nesta perspectiva, os procedimentos didáticos devem privilegiar a construção coletiva dos conhecimentos, mediados pela tecnologia, na qual o professor é um partícipe pró-ativo que intermedeia e orienta esta construção (MORAN, 2007).

Segundo o Censo EAD.BR 2016 (ABED, 2017), os três principais problemas enfrentados pela Educação a Distância no Brasil são a evasão, a resistência dos próprios educadores e a dificuldade de adaptação dos estudantes a essa modalidade de ensino.

2.1 Evasão Escolar

A evasão está entre os temas que tem grande relevância no debate sobre educação no cenário das políticas públicas, sendo um dos temas mais discutidos na educação. Assim, a evasão é um grande desafio para as escolas, pais e para o sistema educacional, sendo ela um dos principais problemas enfrentados pela educação no Brasil (EYNG et al., 2013).

A evasão é definida por Eyng et al. (2013) como um processo que leva o aluno a deixar de frequentar a aula, caracterizando o abandono da escola durante o ano letivo. Segundo os autores, esse processo é gradativo, vai ocorrendo no andamento do curso, e usualmente não é notado.

Queiroz (2001) afirma que a responsabilidade da evasão na educação é tanto do Estado quanto da família. Desta forma, esses dois agentes têm o dever da orientação sócio educacional de crianças e adolescentes. Nessa perspectiva, a Lei de Diretrizes e Bases da Educação (LDB) é bastante clara:

> Art. 2º. A educação, dever da família e do Estado, inspirada nos princípios de liberdade e nos ideais de solidariedade humana, tem por finalidade o pleno desenvolvimento do educando, seu preparo para o exercício da cidadania e sua qualificação ao para o trabalho.

Segundo Manhães et al. (2011), a evasão pode ser definida em três eixos:

- Evasão de curso: o estudante desliga-se do curso em situações diversas: abandono (deixa de se matricular), desistência (oficial), transferência (mudança de curso) ou exclusão por norma institucional.
- Evasão da instituição: o estudante desliga-se da instituição na qual está matriculado.
- Evasão do sistema: o estudante abandona de forma definitiva ou temporária a modalidade de ensino.

Atualmente, a ocupação de uma vaga em uma instituição pública de ensino seguida da evasão tornou-se um problema generalizado, independente da instituição, gerando além de perdas pessoais, também sociais e financeiras. Com isso, a evasão deixa de ser um problema pessoal do estudante e passa a ser um problema que precisa ser combatido com ações do Estado (MANHÃES et al., 2011).

Barroso e Falcão (2004) destacam os fatores que podem desencadear a evasão escolar: (i) Econômicos: impossibilidade de permanecer no curso por questões socioeconômicas; (ii) Vocacionais: o aluno não se identifica com o curso; e (iii) Institucionais: abandono por fracasso nas disciplinas iniciais, inadequação aos métodos de estudo, dificuldades de relacionamento com colegas ou com membros da instituição.

Um relatório desenvolvido pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) classifica o Brasil como a terceira maior taxa de evasão no ensino médio, entre os 100 países com maior Índice de Desenvolvimento Humano (IDH), chegando a alcançar taxas de 24,3% (OLIVEIRA, 2002).

De acordo com Censo EAD.BR 2016 (ABED, 2017) as taxas de evasão reportadas nos cursos a distância são maiores que as nos cursos presenciais. Este censo registrou uma evasão entre 11% e 25% nas instituições que oferecem cursos regulamentados totalmente a distância.

Entre os motivos da evasão investigados pelas instituições e declarados no Censo EAD.BR 2016, a questão financeira é apontada como principal causa de evasão, também aparecem com índices altos a falta de tempo e a falta de adaptação à modalidade. Com isso, um dos desafios na EaD é analisar o desempenho dos alunos e buscar desenvolver metodologias que favoreçam o aprendizado e aumentem o engajamento e a motivação, reduzindo as taxas de evasão.

2.2 Mineração de Dados Educacionais

Segundo Romero e Ventura (2013), a maioria das técnicas tradicionais de mineração de dados, tais como: classificação, agrupamento e técnicas de análise de associação; já foram aplicadas com êxito no domínio da educação. As próximas seções apresentam estas técnicas, seguindo a taxonomia proposta por Baker, Isotani e Carvalho (2011).

2.2.1 Predição

Segundo Baker, Isotani e Carvalho (2011), na predição através da análise e fusão das informações contidas nos dados são criados modelos que podem inferir características e informações sobre esses dados, estas são chamadas de variáveis preditivas (*predicted variables*).

Entre os três tipos de predição, os mais utilizados são classificação e regressão, enquanto que estimação de densidade dificilmente é utilizado devido à dependência estatística dos dados. Os outros dois tipos de predição variam conforme a variável a ser prevista. Quando esta variável é numérica são utilizados geralmente algoritmos de regressão linear e redes neurais. Por sua vez, com variáveis binárias ou categóricas são utilizados algoritmos de classificação como árvores de decisão.

2.2.2 Agrupamento

O agrupamento tem por objetivo a identificação de grupos de dados que apresentem semelhanças entre si em alguma variável ou aspecto, assim auxiliando na tarefa de descoberta de novas informações. Geralmente, os dados são agrupados utilizando alguma medida de distância que decide a semelhança entre os dados, uma vez feita esta etapa, os dados podem voltar a ser analisados, pois podem ser gerados reagrupamentos a partir dos anteriores (ROMERO; VENTURA, 2013).

Na mineração de dados educacionais, o agrupamento pode ser utilizado para agrupar alunos, interações ou até mesmo materiais. Assim, esta etapa passa tanto pela descoberta da relação das

variáveis como pela tentativa de aprendizado de quais variáveis e valores tem um maior impacto sobre alguma outra variável que geralmente é a que se busca prever.

2.2.3 Mineração de Relações

No conceito taxonômico de Baker e Yacef (2009) existem quatro tipos de mineração usadas para identificar relações: regras de associação, correlações, sequências e causas.

A mineração por regras de associação tem como premissa básica a busca por variáveis que tenham seu valor associado ao valor de outras variáveis. Para isso utiliza-se testes condicionais do tipo "se-então" (if-then), assim quando uma variável assume um determinado valor pode-se inferir o valor da segunda variável.

As regras de associação representam padrões existentes em transações armazenadas. Considerando o exemplo de Vasconcelos e Carvalho (2004), a partir de uma base de dados, na qual registram-se os itens adquiridos por clientes, uma estratégia de mineração, com o uso de regras de associação, poderia gerar a seguinte regra: cinto, bolsa!sapato, a qual indica que o cliente que compra cinto e bolsa, com um determinado grau de confiança, compra também sapato. Este grau de certeza de uma regra é definido por dois índices: o fator de suporte e o fator de confiança.

Na mineração de correlações, o objetivo é encontrar correlações lineares entre as variáveis, por exemplo, um conjunto de dados educacionais onde busca-se identificar a nota de um aluno a partir do tempo gasto na aula por esse em tarefas não relacionadas às dadas pelo professor.

Em mineração de sequências, busca-se encontrar uma associação temporal entre os eventos e o impacto deles nas variáveis. Como exemplo, pode-se citar a trajetória dos atos e ações de um determinado aluno e o resultado de sua aprendizagem.

Na mineração de causas, a principal ideia é que se busque identificar eventos que ocasionem outros eventos. Isso ocorre através da análise de padrões de covariância, por exemplo, pode-se citar o comportamento de um aluno em aula. Muitas vezes seu comportamento inadequado está ligado diretamente à sua dificuldade de aprendizagem. Assim, um resultado ruim em determinadas tarefas pode representar na verdade um problema de aprendizagem.

2.2.4 Destilação dos Dados

A destilação dos dados busca facilitar a compreensão de dados complexos e suas características. Essa etapa possibilita que os dados sejam analisados e assim as pessoas os compreendam e identifiquem padrões sobre eles, auxiliando na tomada de decisões. Um exemplo seria a curva de aprendizagem de um aluno, que são representações matemáticas do desempenho de um estudante quando submetido a tarefas de ensino (ARGOTE, 1999). À medida que as repetições são efetuadas, o estudante demanda menos tempo para o aprendizado, seja pela familiaridade adquirida com os meios, seja pela adaptação às ferramentas utilizadas ou pela descoberta de "atalhos" para realização da tarefa (DAR-EL, 2013).

3. TRABALHOS RELACIONADOS

Lykourentzou et al. (2009) propõem um sistema de predição de alunos em situação de risco de evasão que combina os resultados da aplicação de três algoritmos diferentes: redes neurais; máquinas de suporte vetorial e sequência mínima de otimização; e conjunto probabilístico simplificado Fuzzy. Em sua pesquisa, os autores utilizam dados demográficos invariantes no decorrer do curso como sexo e residência, além de dados acadêmicos como performance e nível escolar, e dados variantes como número de interações com o ambiente virtual, notas e a data da entrega dos trabalhos. Com a aplicação dos algoritmos são criados três esquemas diferentes buscando a predição da evasão: (i) um estudante é considerado evadido se pelo menos uma técnica classificou este estudante como tal; (ii) um estudante é considerado evadido se pelo menos duas técnicas indicam essa situação; e (iii) é necessário que as três técnicas classifiquem o aluno como evadido para que este seja assim definido. Os resultados obtidos variam de 73% a 94%, sendo que os mais satisfatórios foram obtidos pelo esquema 1 que chegou a atingir 94%.

A pesquisa de Manhães et al. (2011) tem como objetivo a aplicação de técnica de mineração de dados para identificação precoce de alunos em risco de evasão nos cursos presenciais de graduação em Engenharia da Universidade Federal do Rio de Janeiro. São utilizados dados sobre o desempenho dos alunos em duas disciplinas do primeiro semestre do curso e aplicados 10 diferentes algoritmos para geração dos modelos dentro da ferramenta WEKA (HALL et al., 2009): *OneR* e *JRip* (baseados em aprendizagem de regras), *DecisionTable* (tabela de decisão), *SimpleCart*, J48 e *RandomForest* (árvores de decisão) e *SimpleLogistic* (regressão logística). A acurácia média dos modelos teve variação entre 75% e 80%.

Jayaprakash et al. (2014) propõem a criação de um sistema de alerta de risco quanto ao desempenho do estudante, a fim de diminuir as taxas de evasão e retenção escolares, fornecendo ao aluno um feedback atualizado de seu possível rendimento escolar. Para isto ele utiliza dados demográficos como sexo e idade, interações dos alunos com o ambiente virtual de aprendizagem, desempenho acadêmico anterior, tempo na universidade, tempo on-line no ambiente virtual, dados do teste de aptidão escolar (*Scholastic Assessment Test* (SAT) Verbal e Matemático), entre outros. Assim, são analisados dados de 9.938 alunos, aplicando árvores de decisão com o algoritmo J48, redes *Bayesianas* com o *Naive Bayes*, Máquinas de suporte Vetorial com o SVM/SMO e regressão logística. Na tarefa de predição, todos os algoritmos apresentaram resultados muito próximos, tendo o classificador de regressão logística apresentado resultados ligeiramente maiores que os outros três, com 94,20% de acurácia geral e 66,70% de precisão na predição de alunos em risco de evasão.

Kantorski et al. (2016) propõem uma metodologia visando a predição da evasão em cursos superiores presenciais, aplicada sobre dados de 791 estudantes de um curso de universidade brasileira. Os autores propõem a utilização de dados demográficos como sexo, idade, estado civil, formação básica entre outros e também dados sobre participação em programas de auxílio estudantil e utiliza acompanhamento psicossocial. São aplicados 4 algoritmos com precisão de até 73%.

Burgos et al. (2017) propõem a utilização de mineração de dados e a criação de modelos de predição utilizando regressão logística linear para prever o risco de evasão de alunos. São utilizados dados de 104 alunos de diversos cursos de pequena duração (20 semanas) na modalidade a distância. Os resultados apresentam valores de até 100% de acurácia geral já na quarta semana do curso. Segundo

os autores, a aplicação desta técnica junto a um plano de tutorial diminuiu em 14% o abandono escolar nos cursos em que foram aplicadas.

A análise dos trabalhos relacionados evidencia uma lacuna de pesquisa na área de predição da evasão, pois mesmo com diversos modelos estes costumam levar em conta dados muitos específicos, os quais nem sempre estão disponíveis em diferentes contextos de educação a distância.

Desta forma, este trabalho diferencia-se dos projetos relacionados por apresentar uma proposta de metodologia para geração de modelos de predição de alunos em risco de evasão que utiliza a contagem de interações e suas variações, o que pode permitir contemplar diferentes domínios de aplicação. Assim, os modelos gerados através da metodologia seguida neste trabalho podem ser facilmente generalizados e aplicados em outras instituições que ofereçam cursos na modalidade de ensino a distância.

4. METODOLOGIA

Para o desenvolvimento deste projeto foram utilizados dados de quatro cursos técnicos na modalidade a distância do Câmpus Visconde de Graça (CaVG), do Instituto Federal Sul-rio-grandense (IFSul). Estes cursos são ministrados em dezoito polos localizados em cidades do estado do Rio Grande do Sul. O AVA utilizado por estes cursos é o Moodle (*Modular Object-Oriented Dynamic Learning Environment* – http://moodle.org).

Os cursos são desenvolvidos com atividades semanais, as quais são postadas no AVA pelo professor. Os alunos têm o prazo de uma semana para o desenvolvimento destas atividades, contando com o auxílio dos tutores. Cada curso tem um tempo de realização máximo de 103 semanas, com carga horária total de 1215 horas, divididas pelas disciplinas dentro do período de 24 meses. A situação final do aluno é determinada pelo seu resultado nas avaliações.

O prazo máximo para integralização do curso é de quatro anos, podendo o aluno repetir somente uma vez cada disciplina. Existe a possibilidade de cursar até duas disciplinas como dependência, de forma concomitante às outras disciplinas do curso. Para aprovação o aluno deverá ter média igual ou superior a seis em cada uma das disciplinas da matriz curricular. Considera-se evadido o aluno que passe um período de 365 dias sem interações com o ambiente virtual ou não efetue sua rematrícula anual, sendo nestas situações desligado do curso.

Considerando que este trabalho busca viabilizar a predição dos alunos que estejam em situação de evasão no decorrer do curso, define-se que o aluno será considerado evadido caso abandone, não efetue as atividades no decorrer do curso, bem como não realize sua rematrícula.

Na perspectiva de que os modelos de predição desenvolvidos neste trabalho sejam de fácil generalização para aplicação em outros cursos do IFSul ou mesmo em outras instituições de ensino que utilizem o modelo da Rede e-TEC, optou-se por utilizar as contagens diárias e semanais de interações dos alunos com o ambiente virtual.

Assim, a metodologia definida para o desenvolvimento desse trabalho utiliza a contagem de interações dos estudantes no AVA como a principal informação para a geração dos modelos de predição. As

próximas seções descrevem as características dos dados coletados, o pré-processamento realizado, e a geração e avaliação dos modelos de predição.

4.1 Coleta

Foram coletados os dados registrados em arquivos, correspondentes às interações dos estudantes de cada uma das disciplinas dos quatro cursos técnicos. A Tabela 1 apresenta as quantidades de dados utilizados em cada um destes cursos, assim como a quantidade total de alunos dos mesmos, e as respectivas quantidades e percentuais de alunos concluintes e evadidos.

Tabela 1. Dados Utilizados

Cursos	Quant. Dados	Nº de Alunos	Evadidos (%)	Concluintes(%)
Curso 1	682.773	407	212 (52%)	195 (48%)
Curso 2	1.033.910	729	301 (41,3%)	428 (58,7%)
Curso 3	933.221	615	246 (40%)	369 (60%)
Curso 4	1.051.012	752	354 (47%)	398 (53%)
Totais	3.700.916	2503	1113 (44,5%)	1390 (55,5%)

4.2 Pré-Processamento

O pré-processamento consistiu inicialmente na limpeza dos dados e anonimização dos estudantes. Um sistema em Java foi desenvolvido para realizar o processo de geração de variáveis derivadas e a contagem das interações a partir dos registros. Os dados foram inseridos em um banco de dados, e separados por este sistema por dia e semana em acordo com o calendário dos cursos, considerando datas de início, férias e término dos semestres.

Ao final, as interações foram contabilizadas ao longo das 103 semanas letivas que constituem cursos. Além da contagem de interações semanais (103 semanas), foram contabilizadas também as contagens de interações diárias (721 dias), bem como a média, a mediana e o desvio padrão de interações na semana (103 semanas). A Tabela 2 apresenta as variáveis utilizadas para a geração dos modelos de predição.

Tabela 2. Variáveis Utilizadas

Variável	Descrição
Interações diárias	Contagem de interações diárias (1 até 721 dias)
Interações Semanais	Contagem das interações na semana (1 até 103 semanas)
Média semanal	Média das contagens das interações na semana (1 até 103 semanas)
Mediana semanal	Mediana das contagens das interações na semana (1 até 103)
Desvio padrão semanal	Desvio padrão das contagens das interações na semana (1 até 103)
Situação final no curso	Situação final no curso (normal ou evadido)
Id	Identificação do estudante

4.3 Modelo de Predição baseado em Estatística Descritiva

Como a proposta do trabalho utiliza apenas as contagens de interações e alguns atributos derivados das mesmas, foi desenvolvido um modelo de predição baseado em estatística descritiva para fins de comparação com os modelos de predição gerados por meio de aprendizagem de máquina (vide Seção 4.4).

A premissa é que um modelo simples estatístico baseado em contagens poderia obter desempenho satisfatório, sem a necessidade de utilização de algoritmos de aprendizagem de máquina.

O modelo proposto para comparação utiliza a média e o desvio padrão das interações semanais das turmas e avalia se a média das interações semanais dos estudantes pertence a um determinado intervalo definido pelo desvio padrão. Este modelo considera evadidos aqueles estudantes que apresentam uma média de interações duas vezes abaixo do desvio padrão de interações semanais da turma, bem como aqueles com interações duas vezes acima, pois se observou que também ocorria evasão de estudantes que acessavam o AVA acima da média da turma.

4.4. Geração e Avaliação dos Modelos de Predição

A geração e avaliação dos modelos foi realizada por meio de cinco diferentes algoritmos de aprendizagem de máquina: *Bayes Net*, *Simple Logistic*, *Multilayer Perceptron*, *Random Forest* e *J48*. Para tanto, foi utilizada a biblioteca WEKA integrada ao sistema em Java, desenvolvido para este trabalho. A escolha destes algoritmos se deve a estes estarem entre os mais utilizados, conforme a revisão de literatura dos trabalhos relacionados ao tema (vide Seção 3).

Os modelos foram testados e avaliados em dois cenários diferentes: 1) treinamento e avaliação dentro de um mesmo curso; e 2) treinamento com dados de três cursos e avaliação com os dados do curso restante.

Para o primeiro cenário foi utilizada a técnica de *10-fold cross-validation*, sendo que os modelos são gerados utilizando 9 subconjuntos diferentes e o teste é feito em 1 subconjunto. Esse processo é repetido 10 vezes e a acurácia se dá pela média dos 10 testes.

No segundo cenário, o treino é realizado com dados combinados de 3 dos 4 cursos disponíveis, e o modelo gerado é testado e avaliado no curso restante, por exemplo, treino utilizando dados dos Cursos 1, 2 e 3; e teste e avaliação dos modelos com dados do Curso 4.

A acurácia dos resultados é medida utilizando o percentual de Verdadeiros Positivos (VP), ou seja, os acertos na predição de um estudante evadido sobre a quantidade de evadidos, e no percentual de Verdadeiros Negativos (VN) que corresponde aos acertos na predição de um estudante que irá finalizar o curso sobre a quantidade de estudantes que finalizaram.

5. RESULTADOS

Esta seção apresenta os resultados obtidos pelos modelos em cada um dos cenários citados anteriormente na metodologia do trabalho (vide Seção 4.4), sendo destacados os resultados referentes ao Curso 3.

5.1. Cenário 1: Treinamento e Avaliação dentro de um mesmo Curso

No cenário 1 os modelos são gerados e aplicados na mesma base de dados do curso, utilizando validação cruzada. A Figura 2 apresenta os percentuais de Verdadeiros Positivos (VP) no lado esquerdo e Verdadeiros Negativos (VN) à direita para os modelos gerados com os dados do Curso 3.

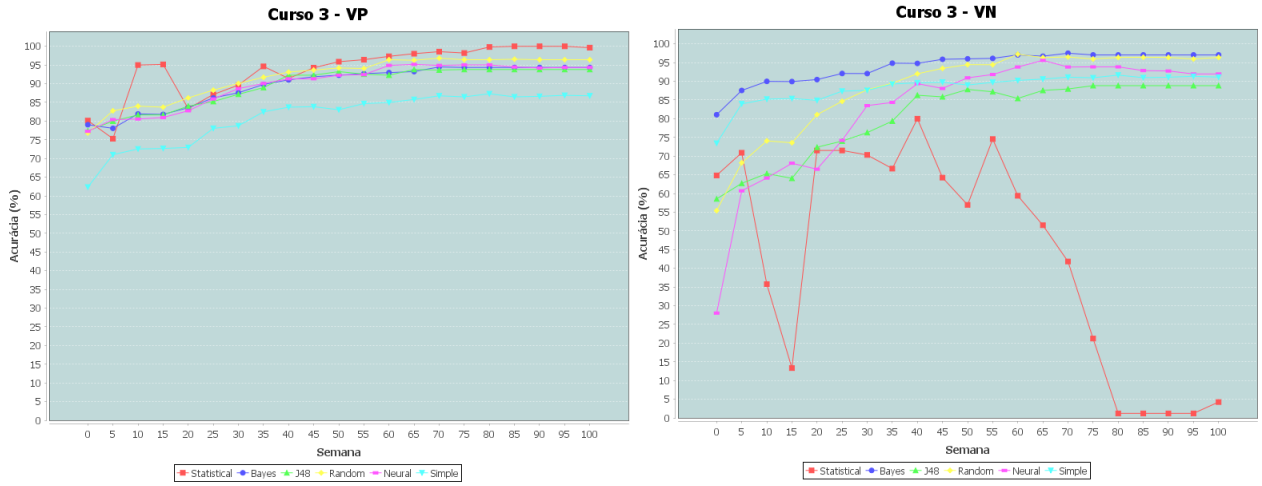

Figura 2. Cenário 1 – Verdadeiros Positivos e Verdadeiros Negativos para o Curso 3

Como é possível observar na Figura 2, de maneira geral os algoritmos apresentam resultados bastante próximos na classificação de estudantes evadidos (VP). Pode-se destacar os resultados obtidos pelo modelo de predição baseado na média e desvio padrão das contagens de interações semanais, além dos resultados do *Random Forest*, que antes da quinta semana de curso obtém as maiores taxas de acerto (ao redor de 83%) dentre os demais algoritmos de aprendizagem de máquina.

Para esse curso, desde a primeira semana as taxas de acerto foram superiores a 75% em quase todos os algoritmos, excetuando-se os resultados obtidos pelo *Simple Logistic*. Com o passar das semanas as taxas de acerto se elevam para 87% antes da semana 25, o que corresponde ao final do primeiro semestre do curso. No segundo semestre do curso já são obtidos resultados próximos a 94% antes do final do mesmo.

Assim, é possível afirmar com quase 95% de exatidão se um aluno irá terminar o curso antes do final do primeiro ano. Nos dois últimos semestres do curso os resultados ultrapassam os 95% com o *Random Forest*, chegando até 96%. Os demais algoritmos têm seus resultados próximos aos 95%. O modelo *Random Forest* se destaca mais ainda quando é avaliada também a acurácia da predição dos casos de estudantes concluintes (VN). Nesse caso, o *Random Forest* apresenta uma acurácia de aproximadamente 85% antes da quinta semana, subindo gradativamente seu desempenho ao longo do tempo e mantendo-se estável até o final do curso.

Ainda, pode-se observar que o desempenho do modelo baseado na média e desvio padrão das contagens atingiu os melhores resultados na predição de estudantes evadidos (VP). Entretanto, esse modelo não é capaz de predizer de maneira satisfatória os casos de alunos concluintes (VN).

5.2. Cenário 2: Treinamento com Dados de Três Cursos e Avaliação com Dados do Curso Restante

Nesse cenário foram utilizados dados de três cursos para treinamento dos modelos, e dados do curso restante para o teste e avaliação dos mesmos. A Figura 3 apresenta os resultados encontrados, utilizando para treino os dados dos Cursos 1, 2 e 4, e para testes os dados do Curso 3.

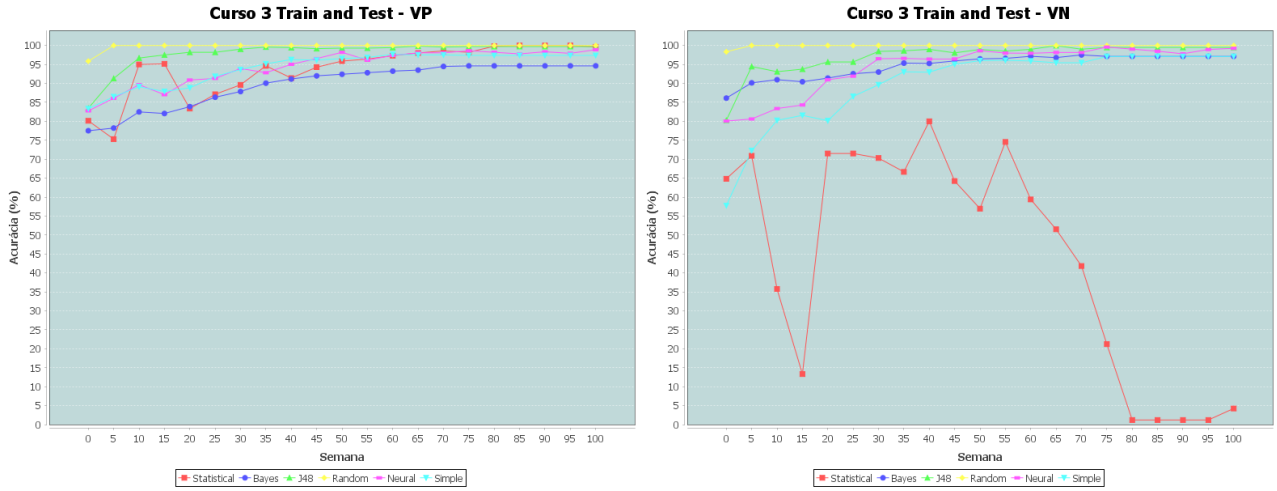

Figura 3. Cenário 2 – Verdadeiros Positivos e Verdadeiros Negativos para o Curso 3

A Figura 3 mostra que desde a primeira semana do curso todos os algoritmos testados obtêm taxas de acurácia na predição de alunos em risco de evasão (VP) superiores a 77%. Novamente, destaca-se o desempenho do algoritmo *Random Forest* que já na primeira semana obteve percentuais acima de 95% para classificação de estudantes evadidos (VP), alcançando 99% a partir da quinta semana de curso e mantendo-se próximo a esse valor até o final do curso.

O algoritmo *J48* também se destaca neste experimento, tendo obtido desde as primeiras semanas 84% de acurácia, com esse valor crescendo gradativamente até alcançar 97% no fim do primeiro semestre do curso (semana 25).

Ainda, na Figura 3 é possível observar os resultados obtidos para classificação de estudantes concluintes (VN). O algoritmo *Random Forest* obteve também as melhores taxas de acurácia. Nos testes feitos com esse algoritmo, desde a primeira semana já são atingidos resultados de 97%, alcançando 99% na semana 5 e 100% nas posteriores até o final do curso. Como pode ser observado nos experimentos apresentados nesse cenário, é possível obter resultados com alto percentual de desempenho tanto na predição de alunos em risco de evasão quanto para a predição de estudantes que tendem a terminar os cursos.

5.3. Discussão dos Resultados

No experimento realizado foi possível observar que os modelos gerados por meio de algoritmos de aprendizagem de máquina apresentam resultados mais estáveis do que o modelo baseado em estatística descritiva (média e desvio padrão) das interações semanais, justificando assim o esforço computacional envolvido na geração dos modelos. É importante ressaltar que os modelos gerados para os demais cursos também apresentam resultados alinhados com os destacados neste artigo referentes ao Curso 3.

Também, os resultados obtidos com ambas as métricas (verdadeiros positivos e verdadeiros negativos) podem ser considerados bastante satisfatórios, permitindo classificar o risco de evasão de estudantes em cursos técnicos a distância de maneira precoce.

A metodologia empregada neste trabalho é resultante de um processo incremental, sendo originada de uma série de experimentos anteriormente realizados pelos autores. Nestes experimentos iniciais,

apenas a contagem de interações semanais era utilizada na geração dos modelos de predição, bem como não havia comparação com um modelo baseado em estatística descritiva.

Assim, o presente trabalho incorpora uma maior granularidade relacionada a contagem das interações (interações diárias), bem como insere atributos derivados das contagens de interações (média, mediana e desvio padrão de interações semanais). Nesse sentido, ao comparar os resultados alcançados neste trabalho com os resultados dos experimentos anteriores, é possível afirmar que uma maior granularidade com relação à contagem das interações, aliada à inserção de atributos derivados permite a geração de modelos de predição mais robustos e com melhor desempenho.

A Tabela 3 apresenta um comparativo entre os melhores resultados obtidos nos experimentos anteriores realizados pelos autores, com os melhores resultados do presente experimento apresentado neste artigo, permitindo verificar uma diferença nos desempenhos dos modelos após a inserção das variáveis anteriormente mencionadas.

Tabela 3. Comparativo entre os Experimentos

Experimento	Semana 1		Semana 25		Semana 50		Semana 75		Semana 100	
	VP	VN	VP	VN	VP	VN	VP	VN	VP	VN
Experimentos anteriores	58	68	82	81	93	93	94	94	97	97
Experimento atual – Cenário 1	80	81	84	92	93	95	97	97	97	97
Experimento atual – Cenário 2	95	97	100	100	100	100	100	100	100	100

Conforme mostra a Tabela 3, desde as primeiras semanas dos cursos já são apresentados resultados satisfatórios na predição de estudantes em risco de evasão.

No comparativo entre os cenários do experimento deste trabalho, podemos notar o impacto que a quantidade de dados na etapa de geração dos modelos tem na predição, uma vez que os resultados do Cenário 2, que utilizou dados de três cursos, apresentaram melhores desempenhos que os resultados do Cenário 1.

6. CONSIDERAÇÕES FINAIS

Neste artigo foram apresentados os resultados obtidos com o emprego de uma metodologia baseada em contagem de interações com o AVA para a predição de estudantes em risco de evasão em cursos técnicos a distância.

Com esta metodologia, o modelo de predição gerado caracteriza-se por uma maior generalização, o que pode viabilizar seu emprego em diferentes instituições, cursos e plataformas, uma vez que utiliza apenas a contagem das interações sem distinção dos tipos de ações realizadas, bem como sem utilizar informações de outras fontes de dados, tais como: dados demográficos, questionários, currículo, cuja disponibilidade pode variar entre plataformas de EaD.

A metodologia foi aplicada em dados de quatro cursos técnicos a distância do CaVG – IFSul. Cabe destacar que este trabalho abrange uma modalidade de ensino que tem sido pouco explorada pela literatura que trata sobre a predição de estudantes em risco de evasão.

Os resultados obtidos podem ser considerados bastante satisfatórios, pois permitem a identificação de estudantes em risco de evasão com elevadas taxas de desempenho antes do final do primeiro semestre dos cursos. Além disso, os experimentos demonstraram que: (i) os modelos gerados por meio de algoritmos de aprendizagem de máquina apresentam melhores desempenhos do que o modelo baseado na média e desvio padrão das interações semanais; e (ii) a inserção de variáveis derivadas com maior granularidade (contagem de interações diárias) melhorou o desempenho do modelo de predição atual em comparação com experimentos anteriores realizados pelos autores.

Na continuidade desta pesquisa, os seguintes temas podem ser considerados em trabalhos futuros: (i) aplicação de uma metodologia de votação que utilize a combinação dos resultados de predição de diferentes modelos; e (ii) criação de um módulo de integração direta do software desenvolvido com a base de dados AVA Moodle.

7. REFERÊNCIAS

ABED, Censo EaD 2016-2017. **Relatório Analítico da Aprendizagem a Distância no Brasil**. Disponível em: abed.org.br/censoead2016/Censo_EAD_2016_portugues.pdf. Acesso em: janeiro de 2018.

ARGOTE, L. **Organizational Learning: Creating, Retaining, and Transferring Knowledge**. 1ª ed. Norwell, MA, USA: Kluwer Academic Publishers, 1999.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v.19, n.02, p.03, 2011.

BAKER, R.; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. **Journal of Educational Data Mining**, v.1, n.1, p.3-16, 2009.

BURGOS, C.; CAMPANARIO, M.; PEÑA, D. de la; LARA, J.; LIZCANO, D.; MARTINEZ, M. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. **Computers & Electrical Engineering**, v.0, p.1-16, 2017.

DAR-EL, E. Human Learning: From Learning Curves to Learning Organizations. **International Series in Operations Research & Management Science**. Springer US, 2013.

DELANO, R.; CORREA, D. Redes na Educação a Distância: Uma Análise Estrutural do Sistema UAB em Minas Gerais. **Revista PRETEXTO**, 2013.

EYNG, A.; GISI, M.; ENS, R.; PACIEVITCH, T. Diversidade e padronização nas políticas educacionais: configurações da convivência escolar. **Ensaio: Avaliação e Políticas Públicas em Educação**, v.21, n.81, p.773-800, 2013.

FARIA, E. T. O professor e as novas tecnologias. ENRICONE, D. (Org.). **Ser Professor**. 4 ed. Porto Alegre: EDIPUCRS, 2004 (p. 57-72).

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v.11, n.1, p.10–18, 2009.

JAYAPRAKASH, S.; MOODY, E.; LAURIA, E.; REGAN, J.; BARON, J. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. **Journal of Learning Analytics**, v.1, n.1, p.6–47, 2014.

KANTORSKI, G.; FLORES, E.; SCHMITT, J.; HOFFMANN, I.; BARBOSA, F. Predição da Evasão em Cursos de Graduação em Instituições Públicas. **Brazilian Symposium on Computers in Education**, v.27, n.1, p.906, 2016.

LYKOURENTZOU, I.; GIANNOUKOS, I.; NIKOLOPOULOS, V.; MPARDIS, G.; LOUMOS, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. **Computers & Education**, v.53, n.3, p.950–965, 2009.

MANHÃES, L.; CRUZ, S.; COSTA, R.; ZAVALETA, J.; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. **Anais do XXII Simpósio Brasileiro de Informática na Educação**, Aracaju, 2011.

MORAN, J. **A educação que desejamos: novos desafios e como chegar lá**. Campinas: Papirus, 2007.

OLIVEIRA, G. Uma discussão sobre o conceito de desenvolvimento. **Revista da FAE**, Curitiba, v.5, n.2, p.37-48, maio/ago. 2002.

QUEIROZ, L. **Um estudo sobre a evasão escolar: para se pensar na inclusão escolar**. Associação Nacional de Pós-Graduação e Pesquisa em Educação (Anpad). Disponível em: www. anped. org. br/reunioes/25/. Acesso em: janeiro de 2018.

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

VASCONCELOS, L.; CARVALHO, C. **Aplicação de regras de associação para mineração de dados na web**. Instituto de Informática da Universidade Federal de Goiás, Relatório Técnico, novembro de 2004.

Agradecimentos

Esse trabalho foi financiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) por meio do Edital Universal 01/2016 processo 404369/2016-2.