Predição de estudantes com risco de evasão em cursos técnicos a distância

Emanuel Marques Queiroga[1], **Cristian Cechinel**[1,2], **Ricardo Matsumura de Araújo**[1]

[1]Programa de Pós-Graduação em Computação
Centro de Desenvolvimento Tecnológico (CDTec)
Universidade Federal de Pelotas (UFPel)
Pelotas, RS, Brasil

[2]Centro de Ciências, Tecnologias e Saúde (CTS)
Universidade Federal de Santa Catarina (UFSC)
Araranguá, SC, Brasil

{emanuel.queiroga,ricardo}@inf.ufpel.edu.br

contato@cristiancechinel.pro.br

Abstract. *The present paper describes an approach for detecting possible dropout students in technical distance learning courses. The proposed method uses only the count of students interactions inside a Learning Management System (LMS), along with other attributes derived from the counts. Such strategy allows better generalization in different platforms and LMS, as it does not rely on the differences among the interaction types, or use other darta sources besides LMS logs. Predictive models were trained and tested with data from 4 technical distance learning courses in two different scenarios: 1) train and test with data from one course, and 2) train with data combined from 3 courses and test with data of the remaining course. Results point out it is possible to predict dropout students in the first weeks of the courses with average accuracy rates of 75% in most of scenarios, achieving 95% in the best case scenarios.*

Resumo. *O presente trabalho apresenta uma abordagem para a detecção de alunos em risco de evasão em cursos técnicos a distância que utiliza apenas a contagem de interações dos estudantes dentro do AVA, além de atributos derivados dessas contagens. A premissa inicial é de que essa estratégia permite uma maior generalização em diferentes plataformas e AVA, uma vez que não utiliza diferenciações entre os tipos de interações, nem informações de outra ordem encontradas fora do AVA (dados demográficos, exames, questionários, etc). Os modelos de predição foram testados e treinados com dados de 4 diferentes cursos técnicos EAD em dois cenários diferentes: 1) treino e teste com dados de um mesmo curso, e 2) treino com dados de 3 cursos e teste com dados do curso restante. Os resultados apontam a possibilidade de predição de estudantes em risco de evasão já nas primeiras semanas dos cursos com taxas de desempenho próximas a 75% na maioria dos cenários, e chegando a 95% nos melhores casos.*

1. Introdução

A Educação a Distância (EAD) apresenta uma alternativa para o acesso aos programas de qualificação profissional no Brasil. Entretanto, junto com a utilização da EAD surgiram

DOI: 10.5753/cbie.sbie.2017.1547

problemas como os altos percentuais de evasão. O Ambiente Virtual de Aprendizagem (AVA) é o "local virtual" onde os cursos na modalidade a distância (e semipresenciais) normalmente ocorrem. Uma das plataformas mais utilizadas atualmente no Brasil é o Moodle, que oferece áreas para apresentação de conteúdos em diversos formatos, além de atividades de verificação da aprendizagem e espaços para interação síncrona (*chats*) e assíncrona (fóruns). A organização do AVA de maneira apropriada permite ao aluno e ao professor um acompanhamento organizado e sistematizado daquilo que deve ser estudado a cada semana ao longo de toda uma disciplina. O AVA também mantém registro em seus *logs* de todas as interações virtuais que ocorrem ao longo do processo de ensino-aprendizagem.

Um dos principais desafios da EAD é obter a diminuição dos atuais índices de evasão, que conforme o Censo EAD (CensoEAD, 2013), foi de 18,6% em 2010, 20,5% em 2011, 11,74% em 2012 e 16,94% em 2013 nos cursos autorizados pelo Ministério da Educação (MEC). Para ilustrar, em 2013 haviam 5754 cursos autorizados pelo MEC e a taxa de matrículas anual foi de 882.843, sendo que o conjunto total de alunos evadidos foi de aproximadamente 150.000.

Segundo [Manhães et al. 2011], a detecção precoce de grupos de alunos com risco de evasão é uma condição importante para reduzir o problema da evasão pois possibilitaria proporcionar algum tipo de atendimento direcionado a situação específica de cada aluno. Ainda segundo [Manhães et al. 2011], atualmente o processo de identificação desse grupo de alunos é manual, subjetivo, empírico e sujeito a falhas, dependendo primordialmente da experiência acadêmica e do envolvimento dos docentes. Considerando que os docentes desempenham inúmeras atividades, assim como também a grande quantidade de alunos normalmente existente em cursos EAD, é bastante difícil acompanhar e reconhecer as necessidades de cada aluno e identificar aqueles alunos que apresentam risco de evasão. A mineração de dados surge como uma alternativa para o tratamento e a descoberta de conhecimento dentro desse grande volume de dados gerados pelos AVA. Atualmente a EDM vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino [Baker et al. 2011]. Essa área pode ajudar as instituições a criarem modelos de predição que tenham condições de avaliar as chances de um determinado acadêmico evadir. A aplicação da EDM pode possibilitar o tratamento diferenciado entre os alunos, dedicando formas de auxílio diferenciadas a um determinado aluno que esteja com uma probabilidade maior de evasão.

O presente trabalho apresenta uma abordagem para a detecção de alunos em risco de evasão em cursos técnicos a distância. Para isso, a metodologia utilizada considera apenas a contagem de interações dos estudantes dentro do AVA e atributos derivados dessas contagens. A premissa inicial é de que essa estratégia permite uma maior generalização em diferentes plataformas e AVA, uma vez que não utiliza diferenciações entre os diferentes tipos de interações, nem informações de outra ordem encontradas fora do AVA (dados demográficos, etc). O artigo está estruturado da seguinte maneira. A Seção 2 apresenta alguns trabalhos relacionados com o problema de predição de estudantes em risco. Na Seção 3 são descritos os dados e o método utilizado nos experimentos realizados. A Seção 4 discute os resultados alcançados, e a Seção 5 apresenta as conclusões do trabalho, assim como sugestões de trabalhos futuros.

2. Trabalhos Relacionados

Diversos trabalhos buscam modelar o comportamento de estudantes na EAD, utilizando uma variedade de técnicas que podem se diferenciar em sua finalidade como a predição dos resultados que o estudante obterá nas avaliações, a predição do risco de evasão dos estudantes, ou o agrupamento dos mesmos por similaridade [Romero and Ventura 2013]. Por exemplo, [Lykourentzou et al. 2009] propôs um sistema de predição de alunos em situação de risco de evasão que combine os resultados da aplicação de 3 algoritmos diferentes (Redes Neurais, Support Vector Machine, e sequência mínima de otimização (SVM/SMO) combinado com conjunto probabilístico simplificado Fuzzy ARTMAP - PESFAM). Em sua pesquisa o autor utiliza 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 até mesmo a data da entrega dos trabalhos. Com a aplicação dos algoritmos são criados 3 esquemas diferentes buscando a predição da evasão, sendo eles: 1) um estudante é considerado evadido se pelo menos uma técnica classificou este estudante como tal, 2) um estudante é considerado evadido se pelo menos duas técnicas indicam essa situação e, 3) um estudante é considerado evadido quando as 3 técnicas classifiquem o aluno como evadido. Os resultados obtidos variaram de 73% a 94%, sendo que os mais satisfatórios foram obtidos pelo esquema 1 que chegou a atingir 94%.

[Jayaprakash et al. 2014] desenvolveu um sistema de alerta de risco relacionado ao desempenho do aluno com o objetivo de diminuir as taxas de evasão e retenção escolares. O sistema fornece ao aluno um feedback atualizado de seu possível rendimento escolar. Para isto o sistema utiliza diferentes tipos de dados, tais como: demográficos (sexo e idade), interações dos alunos com o AVA, desempenhos acadêmicos anteriores, tempo na universidade, tempo online no AVA, e resultados do teste de aptidão escolar (SAT Verbal e Matemático). Foram gerados distintos modelos de predição (J48, redes Bayesianas com o Naybe Bayes, Máquinas de suporte Vetorial com SVM/SMO e regressão logística) com dados de 9938 alunos. Os classificadores apresentaram resultados muito próximos, sendo que o classificador de regressão logística apresentou resultados ligeiramente maiores que os demais com 94,20% de acurácia geral e 66,70% de precisão na predição de alunos em risco de evasão.

Ainda, a pesquisa de [Manhães et al. 2011] aplica técnicas de mineração de dados para identificação precoce de alunos em risco de evasão nos cursos de graduação em Engenharia presencial da Universidade Federal do Rio de Janeiro. Os autores utilizaram dados sobre o desempenho dos alunos em duas disciplinas do primeiro semestre do curso, e aplicaram 10 diferentes algoritmos para geração dos modelos obtendo uma acurácia média entre 75% e 80%. [Halawa et al. 2014] propõe um preditor de alunos em risco de evasão que antecipe a situação em 14 dias a partir dos dados das interações dos alunos com o ambiente. Dentre as informações utilizadas no processo de predição estão: se o aluno assistiu todas as vídeo aulas, se ignorou algum determinado material ou atividade, o atraso do aluno no acompanhamento do material(se ele está em dia com as aulas ou tem aulas atrasadas) e o seu desempenho nas atividades. Os alunos são então classificados em três bandeiras (verde - baixo risco de evasão, amarela - risco de evasão moderado, vermelha - alto risco de evasão). Os autores não reportam quais os tipos de classificadores utilizados mas relatam uma acurácia na predição entre 40% e 50% com duas semanas de antecedência ao evento de evasão. Por último [Burgos et al. 2017] utiliza regressão

logística linear para prever o risco de evasão de alunos em um curso de pequena duração na modalidade a distância. Os autores utilizaram dados de 104 alunos e geraram um modelo capaz de predizer com até 100% de acurácia geral as chances de evasão dos alunos já na quarta semana do curso. De acordo com os autores, a aplicação dessa técnica aliada a um plano tutorial ajudou a reduzir em 14% a evasão escolar nos cursos alvo.

3. Metodologia

A metodologia seguida para o desenvolvimento desse trabalho é baseada em trabalhos anteriores [Queiroga et al. 2016] [Detoni et al. 2015] e utiliza a contagem de interações dos estudantes no AVA como a principal informação para a geração dos modelos de predição. As subseções a seguir descrevem as características dos dados coletados, o pré-processamento realizado, e a geração e avaliação dos modelos de predição.

3.1. Coleta

Foram coletados os logs de interações de cada uma das disciplinas de 4 cursos técnicos ministrados no Instituto Federal Sul-rio-grandense (IFSUL). A tabela 1 apresenta as quantidade de logs em cada um dos cursos incluídos no trabalho, assim como a quantidade total de alunos desses cursos, e as respectivas quantidades e percentuais de alunos concluintes e evadidos.

Tabela 1. Dados utilizados

Cursos	Quant. Logs	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%)

3.2. Pré-processamento dos dados

O pré-processamento consistiu inicialmente na limpeza dos dados e anonimização dos acadêmicos. Um sistema em Java foi desenvolvido para facilitar o processo de geração de variáveis derivadas e a contagem das interações a partir dos logs. Os dados foram inseridos em um banco de dados, e separados pelo 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 de 103 semanas letivas que compunham os cursos. Além da contagem de interações semanais (103 semanas), foram contabilizadas também as contagens de interações diárias (721 dias), e a média, mediana e desvio padrão de interações na semana (103 semanas). A Tabela 2 apresenta as variáveis utilizadas para a geração dos modelos de predição.

3.3. Modelo de predição baseado na média e desvio padrão das interações

Como a proposta do trabalho utiliza apenas as contagens de interações e alguns atributos derivados das mesmas, desenvolveu-se também um modelo de predição baseado em estatísticas descritivas para fins de comparação com os modelos de predição gerados por meio de aprendizagem de máquina. A premissa inicial era de que um modelo simples

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	Id do estudante

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 pertencem ou não a um determinado intervalo definido pelo desvio padrão. O modelo proposto considera evadidos aqueles estudantes que apresentam uma média de interações duas vezes acima ou duas vezes abaixo do desvio padrão de interações semanais da turma (observou-se que acadêmicos que acessavam o AVA muito além da média da turma também evadiam). Esse modelo foi incluído nas análises e resultados que serão apresentados a seguir.

3.4. Geração e avalição dos modelos de predição

A geração e avaliação dos modelos foi realizada por meio de 5 diferentes algoritmos (Bayes Net, Simple Logistic, Multilayer Perceptron, Random Forest e J48) e utilizando a biblioteca WEKA integrada ao software desenvolvido para o trabalho. A escolha destes algoritmos se deve a estes serem alguns dos mais utilizados em pesquisas relacionadas ao tema e por já terem apresentado resultados satisfatórios em testes iniciais realizados [ref blind]. Os modelos foram testados e avaliados em 2 cenários diferentes, sendo eles: 1) treino e avaliação dentro de um mesmo curso, 2) treino com dados de 3 cursos e avaliação com os dados do curso restante. Para o primeiro cenário foi utilizada a técnica de avaliação cruzada com 10 pastas (10-fold crossvalidation), 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 (ex: 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) (Acertos na predição de um estudante evadido sobre a quantidade de evadidos) e no percentual de Verdadeiros Negativos (VN) (Acertos na predição de um estudante que irá finalizar o curso sobre a quantidade de estudantes que finalizaram).

4. Resultados

Esta seção apresenta os resultados obtidos pelos modelos em cada um dos cenários definidos anteriormente na metodologia do trabalho. Por restrições de espaço, são apresentados apenas os resultados em cada cenário para o Curso 3.

4.1. Cenário 1 - Treino 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 1 apresenta os percentuais de verdadeiros positivos (VP) (esquerda) e verdadeiros negativos (VN) (direita) para os modelos gerados com os dados do Curso 3.

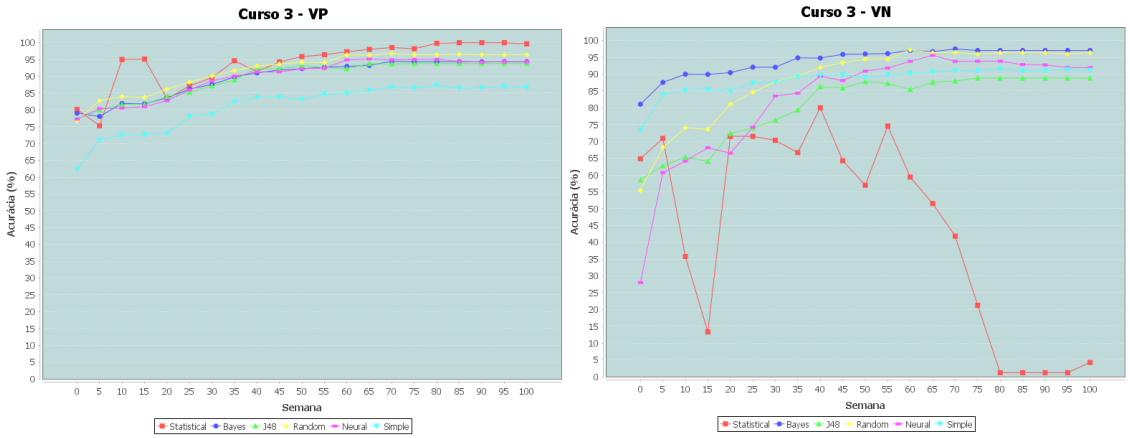

Figura 1. Cenário 1 - Verdadeiros Positivos (esquerda) e Verdadeiros Negativos (direita) para o Curso 3

Como é possível observar na Figura 1, de maneira geral os algoritmos apresentam resultados bastante próximos na classificação de estudantes evadidos (resultados a esquerda). 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 (Curso 3), já 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 (que seria o fim 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 dizer com quase 95% de exatidão se um aluno irá terminar o curso antes do final do primeiro ano do mesmo. Nos dois últimos semestres do curso os resultados ultrapassam os 95% com o Random Forest, chegando até 96%. Os demais algoritmos tem seus resultados próximos aos 95%. O modelo Random Forest se destaca mais ainda quando avaliamos também a acurácia da predição dos casos de estudantes concluintes - verdadeiros negativos (figura da direita). 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. Como pode ser observado, ainda que o desempenho do modelo baseado na média e desvio padrão das contagens tenha atingido os melhores resultados na predição de estudantes evadidos (figura da esquerda), esse modelo não é capaz de predizer de maneira satisfatória os casos de alunos concluintes (figura da direita). De fato, esse experimento especificamente (cenário 1 - curso 3) foi o único onde este modelo obteve resultados próximos aos demais modelos para a predição de acadêmicos evadidos.

4.2. Cenário 2 - Treino com dados de 3 cursos e avaliação com dados do curso restante

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

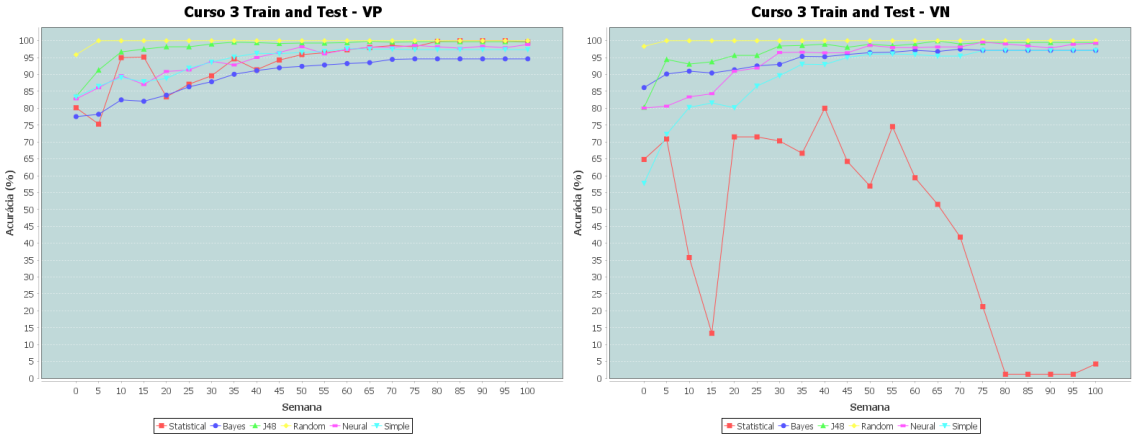

Figura 2. Cenário 2 - Verdadeiros Positivos (esquerda) e Verdadeiros Negativos (direita) para o Curso 3

Como é possível observar na Figura 2 , desde a primeira semana de curso todos algoritmos testados obtém taxas de acurácia na predição de alunos em risco de evasão superiores a 77%. Novamente se destaca o desempenho do algoritmo Random Forest que já na primeira semana obteve percentuais acima de 95% para classificação de estudantes evadidos (figura da esquerda), 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 merece destaque 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).

Na Figura 2 (direita) é possível observar os resultados obtidos para classificação de estudantes concluintes (VN). Novamente o algoritmo Random Forest obteve 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 posteriores 100% 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. Ainda, os modelos gerados por meio de algoritmos de aprendizagem de máquina apresentam resultados mais estáveis do que o modelo baseado simplesmente na média e desvio padrão das interações semanais, justificando assim o esforço computacional envolvido na geração dos modelos. É importante ressaltar que aqui estamos apenas apresentando os resultados para os modelos gerados para o Curso 3, sendo que os modelos gerados para os demais cursos também apresentam resultados satisfatórios.

4.3. Discussão dos Resultados

De modo geral, os resultados obtidos em ambas as métricas (verdadeiros positivos e verdadeiros negativos) foram bastante satisfatórios, permitindo classificar o risco de evasão de estudantes em cursos técnicos a distância de maneira precoce. Os dados utilizados estavam ”naturalmente”balanceados, com percentuais de estudantes evadidos e concluintes relativamente parecidos em cada conjunto. Ainda que não tenha sido possível apresentar os resultados de todos os experimentos realizados, cabe ressaltar que os melhores desempenhos foram encontrados para modelos gerados por meio das bases de dados mais balanceadas.

A metodologia atual empregada é resultado de um processo incremental de uma série de experimentos anteriormente realizados[ref blind][ref blind]. Nos experimentos iniciais, apenas a contagem de interações semanais era utilizada na geração dos modelos de predição, e não havia comparação com nenhum modelo baseado em estatísticas descritivas básicas. O presente trabalho incorpora uma maior granularidade relacionada a contagem das interações (contagem das interações diárias), assim como também insere alguns atributos derivados das contagens de interações (média, mediana e desvio padrão de interações semanais). Nesse sentido, ao comparar os resultados alcançados nesse trabalho com os resultados dos trabalhos anteriores, é possível afirmar que uma maior granularidade com relação a contagem das interações (contagens de interações diárias) aliada a inserção de atributos derivados permite a geração de modelos de predição mais robustos e com melhor desempenho.

A Tabela 3 apresenta os melhores resultados obtidos nos experimentos anteriores com os melhores resultados do presente experimento, permitindo verificar uma diferença considerável nos desempenhos dos modelos após a inserção das variáveis anteriormente mencionadas.

Tabela 3. Comparativo entre Experimentos dos Autores

Experimento	Semana 1		Semana 25		Semana 50		Semana 75		Semana 100	
	VP	VN	VP	VN	VP	VN	VP	VN	VP	VN
Experimento Trabalhos Anteriores [ref blind]	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

Como é possível observar, desde as primeiras semanas dos cursos já são apresentados resultados satisfatórios na predição de estudantes em risco de evasão. No comparativo direto entre os experimentos 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 utilizaram dados de 3 cursos) apresentaram melhores desempenhos que os resultados do Cenário 1.

5. Conclusões

O presente artigo apresentou os resultados de uma abordagem para a predição de estudantes em risco de evasão que utiliza contagens de interações dos mesmos dentro do AVA. A proposta tem como premissa permitir uma maior generalização no momento de replicar a metodologia em outros cursos e plataformas, uma vez que utiliza apenas a contagem das interações dentro do AVA sem distinção dos tipos de ações realizadas, e sem utilizar informações de outras fontes de dados (dados demográficos, questionários, currículo, etc), cuja disponibilidade podem variar entre plataformas de EAD.

A abordagem foi aplicada em dados reais de 4 cursos técnicos a distância, sendo estes um tipo de curso que normalmente não aparecem na literatura sobre a predição de estudantes em risco. Os resultados encontrados podem ser considerados bastante satisfatórios já que permitem a identificação de estudantes em risco de evasão com boas taxas de desempenho antes mesmo do final do primeiro semestre dos cursos. Ainda, os experimentos demonstraram que: 1) os modelos gerados por meio de algoritmos de aprendizagem de máquina apresentam melhores desempenhos do que o modelo baseado em médias e desvios padrões das interações semanais, e 2) que a inserção de variáveis derivadas com maior granularidade (contagem de interações diárias) ajudou a melhorar o desempenho dos modelos em comparação a experimentos anteriores.

Trabalhos futuros incluem a aplicação de uma metodologia de votação (*ensemble*) que utilize a combinação dos resultados de predição de diferentes modelos, além da criação de um módulo de integração direta do software desenvolvido com a base de dados do Moodle institucional.

6. Agradecimentos

O presente trabalho foi realizado com apoio do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) por meio do Edital Universal 01/2016 processo 404369/2016-2 (Projeto: “Comparação entre Diferentes Abordagens na Modelagem e Identificação de Acadêmicos em Risco em Cursos de Educação a Distância”), e do Instituto Federal Sul-rio-grandense (Projeto: ”Aplicação de Mineração de Dados na Predição de Evasão de Alunos da Educação a Distância”).

Referências

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., and Martínez, M. A. (2017). Data mining for modeling students’ performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 0:1–16.

Detoni, D., Cechinel, C., and Matsumura Araújo, R. (2015). Modelagem e predição de reprovação de acadêmicos de cursos de educação a distância a partir da contagem de interações. *Revista Brasileira de Informática na Educação*, 23(3).

Halawa, S., Greene, D., and Mitchell, J. (2014). Dropout Prediction in MOOCs using Learner Activity Features. *European MOOC Summit, EMOOCs*, 37(March):1–10.

Jayaprakash, S. M., Moody, E. W., Lauria, E. J. M., Regan, J. R., and Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, 1(1):6–47.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., and Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3):950–965.

Manhães, L. M. B., Cruz, S. d., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *Anais do XXII SBIE-XVII WIE, Aracaju*.

Queiroga, E., Cechinel, C., Araújo, R., and da Costa Bretanha, G. (2016). Generating models to predict at-risk students in technical e-learning courses. In *Learning Objects and Technology (LACLO), Latin American Conference on*, pages 1–8. IEEE.

Romero and Ventura (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.