Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações

Douglas Detoni e Ricardo Matsumura Araujo
Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas
Rua Gomes Carneiro, 1 – Pelotas – RS - Brasil
douglasdetoni92@gmail.com, ricardo@inf.ufpel.edu.br

Cristian Cechinel
Faculdade de Educação – Universidade Federal de Pelotas
Rua Benjamin Constant, 897 – Pelotas – RS – Brasil
contato@cristiancechinel.pro.br

Abstract. *The possibility of predicting the risk of an student failing a course is useful for tutors and teachers, as this allows them to change their methods to avoid failures. In this paper we show results of attempts to train machine learning models to perform this task, using only interaction counts as attributes. We show that bayesian networks are adequate for the problem and that introducing attributes derived from the counts (e.g. means) are useful for more accurate predictions when data is sparse. We also show the possibility of training models in different source of examples, such as between different groups and semesters.*

Resumo. *A possibilidade de prever com antecedência o risco de reprovação de um estudante em um curso a distância é de grande valia para professores e tutores, que podem adequar seus métodos para evitar a reprovação. Neste trabalho mostramos resultados da aplicação de técnicas de aprendizado de máquina nesta tarefa, utilizando como atributos unicamente contagens de interações ao longo do tempo. Mostramos que redes bayesiana são adequadas ao problema e que a introdução de atributos derivados das contagens (e.g. médias) são úteis para previsões mais precisas quando a quantidade de dados é esparsa. Mostramos ainda a possibilidade de treinar os modelos em diferentes situações, como entre turmas e entre semestres diferentes.*

1. Introdução

A Educação a Distância (EAD) no Brasil tem se consolidado, com diversos estudantes optando por essa modalidade de ensino para ampliar suas formações e realização profissional. Em 2012, segundo dados Censo EAD (CensoEaD, 2013), foram ofertados 9.376 cursos. Um total de 5.772.466 matrículas foram registradas, 52,5% de aumento em relação a 2011. Já o número de conclusões, em 2012, foi de 1.589.374. Esses dados mostram a evolução da Educação a Distância e como ela está se tornando uma ferramenta muito importante na formação dos cidadãos brasileiros. Contudo, a EAD ainda enfrenta alguns obstáculos a serem ultrapassados. Resistência de educandos e educadores, desafios organizacionais e custos de produção são alguns deles mas, sem

dúvida, um dos maiores obstáculos é a evasão de alunos dos cursos e instituições. No ano de 2012, a evasão foi de 11,74% nos cursos autorizados segundo dados do Censo EAD.

Um dos principais diferenciais de cursos de EAD é a grande quantidade de dados gerada pelas interações no ambiente educacional, o que abre novas possibilidades para estudar e compreender estas interações (Hämäläinen e Vinni, 2011; Mishra et al., 2014). Assim, algumas áreas de pesquisas surgiram nos últimos anos com intuito de auxiliar em questões como essas. A *Educational Data Mining* (EDM) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no contexto educacional (Romero & Ventura, 2010). Juntamente com a EDM temos a *Learning Analytics* (LA), outra área de pesquisa emergente. Segundo definição criada na conferência *Learning Analytics and Knowledge* (LAK) em 2012, LA busca medir, coletar, analisar e reportar dados sobre estudantes e seus contextos, com o propósito de entender e otimizar seu aprendizado e o ambiente onde ele ocorre (Brown, 2012). A interação dos alunos com os ambientes virtuais de aprendizagem (AVA) proveem os dados que alimentam as pesquisas na área e possibilitam a descoberta de novos conhecimentos.

Um dos grandes desafios dos pesquisadores é desenvolver métodos capazes de prever o comportamento dos estudantes, de modo a possibilitar a intervenção de professores/tutores, ou demais envolvidos, visando resgatar o estudante antes que ele reprove (Macfadyen e Dawson, 2010). Outros objetivos são detectar estudantes que precisam de ajuda, prever o desempenho, classificar diferentes grupos de estudantes, detectar trapaças no sistema, entre outros (Romero & Ventura, 2010). Segundo Romero and Ventura (2010), existe a necessidade de desenvolver ferramentas específicas e fáceis de usar, para que professores/tutores não familiarizados com as técnicas de EDM e LA também possam se valer das descobertas das áreas. É importante também integrar as ferramentas a AVAs tradicionalmente já usados. Outro aspecto a ser trabalhado é a padronização dos dados e modelos. Muitas ferramentas possuem padrões próprios e específicos, os quais não podem ser integrados por outras ferramentas.

O principal objetivo deste trabalho é desenvolver e testar técnicas e modelos de aprendizado de máquina na predição da reprovação de alunos de educação a distância utilizando o registro de interações entre alunos, tutores e professores em cursos da Universidade Federal de Pelotas no ambiente Moodle.

Adicionalmente, é nosso objetivo analisar unicamente atributos agregados - i.e. números totais de interações, sem especificar o tipo de atividade sendo realizada pelo estudante. Este objetivo é motivado pela heterogeneidade de atividades entre cursos e plataformas; ao utilizar como atributo atividades específicas (e.g. fórum, *chat*), acaba-se por ancorar a técnica aos cursos específicos sendo analisados, tornando-a pouco generalizável.

Este trabalho está estruturado da seguinte maneira. A seção 2 descreve alguns trabalhos relacionados com a predição da situação de acadêmicos em AVAs que se aproximam do presente trabalho. A seção 3 apresenta o contexto de coleta dos dados e o pré-processamento realizado. Na seção 4 são descritos as análises realizadas e na seção 5 os resultados encontrados. As considerações finais e propostas de trabalhos futuros são apresentadas na seção 6.

2. Trabalhos Relacionados

A previsão do desempenho, ou situação, acadêmica de alunos é um tema muito estudado entre pesquisadores. Manhaes, da Cruz, Costa, Zavaleta, and Zimbrao (2011) utilizaram técnicas de mineração de dados para prever a evasão de estudantes em cursos presenciais da Escola Politécnica da Universidade do Rio de Janeiro. Dez modelos diferentes foram testados, com acurácia média variando entre 75% e 80%, com Perceptron de Múltiplas Camadas e Florestas Aleatórias apresentando os melhores desempenhos.

Gottardo, Kaester, and Noronha (2012), apresentam resultados de tentativas de prever a avaliação de desempenho de alunos de educação a distância. Os autores relatam taxas de precisão acima de 74%, com o uso de uma grande variedade de atributos. O presente trabalho diferencia-se ao tentar prever a reprovação de alunos e não seus desempenhos específicos, além de prover uma análise mais detalhada utilizando unicamente atividades agregadas.

Rodrigues, de Medeiros, and Gomes (2013) apresentaram um estudo da viabilidade do uso de um modelo de regressão linear para também prever o desempenho dos alunos. Este trabalho também utilizou dados do ambiente virtual de aprendizagem Moodle, além da utilização de séries temporais. Diversos atributos foram utilizados, incluindo a quantidade de interações ao longo das semanas em fóruns, vídeos, materiais de apoio, entre outros.

Gotardo et al. (2013) realizam a tarefa de predição do desempenho do aluno utilizando Sistemas de Recomendação e Acoplamento de Classificadores. Os dados são relativos a um curso com 252 alunos. Os dados para treinamento dos classificadores foram extraídos da tabela de logs e separados por tipos, como por exemplo: assignment, forums, upload, entre outros. Cada tipo transformou-se em um atributo, juntamente com a situação acadêmica do aluno. Os classificadores J48 e *Naive Bayes* foram utilizados no trabalho. O melhor resultado foi aproximadamente 80% de predição, utilizando 40% do conjunto de dados, com o classificador *Naive Bayes*.

3. Descrição dos Dados

3.1. Contexto dos cursos

A Universidade Federal de Pelotas (UFPEL) oferece cursos de graduação, pós-graduação e extensão na modalidade a distância. São mais de 4000 alunos matriculados e espalhados por diversos polos do Rio Grande do Sul.

Para a realização do trabalho, foram obtidos os dados anonimizados dos cursos a distância de Licenciatura Educação do Campo (CLEC) e de Licenciatura em Pedagogia (CLPD). É importante notar que estes cursos seguem um modelo de execução das disciplinas que difere do modelo presencial convencional. Normalmente, um curso é dividido em semestres, onde diversas disciplinas são ministradas paralelamente ao longo do mesmo. No caso do CLEC e do CLPD, as disciplinas são executadas de maneira sequencial (com no máximo 2 disciplinas acontecendo em paralelo). Normalmente são executadas em torno de 5 disciplinas por semestre com duração média de cinco a sete semanas.

Foram utilizadas disciplinas do primeiro e segundo semestre dos cursos citados. Ambos os semestres ocorreram no ano de 2013, totalizando 329 estudantes únicos. Contabilizando os 4 semestres utilizados, temos um total de 604 alunos, 72 tutores e 80 professores. A Tabela 1 apresenta o número de alunos para cada situação.

Tabela 1. Total de alunos em cada curso e semestre

Curo	Semestre	Alunos	Aprovados	Reprovados
Educação do Campo (CLEC)	1º	133	103	30
	2º	94	88	6
Pedagogia (CLPD)	1º	196	137	59
	2º	172	119	53

3.2 Coleta e Pré-processamento dos Dados

Os dados foram extraídos dos registros de acesso dos cursos fornecidos pela plataforma Moodle. Destes registros, para cada interação de cada usuário, extraiu-se apenas o identificador do usuário, sua função (professor, tutor ou aluno) e a data e hora da interação. Assim, ao contrário de trabalhos anteriores, procuramos utilizar apenas o número total de interações dos estudantes, tutores e professores com o sistema, sem detalhamento do tipo de atividade sendo realizada. Isto permite que os classificadores gerados sejam aplicáveis a diferentes cursos, plataformas e execuções, sem gerar uma dependência em atributos excessivamente específicos. No total foram extraídas cerca de 835 mil interações.

Cada aluno foi rotulado como *aprovado* ou *reprovado* em cada disciplina de acordo com a situação informada nos registros acadêmicos. Cada disciplina possui duração de 7 semanas e as interações de cada aluno foram agrupadas de forma semanal. Assim, cada aluno em uma disciplina é descrito por um vetor de 7 valores indicando o número total de interações daquele aluno em cada semana.

Considerando que o objetivo principal é predizer a situação final do aluno com a maior antecedência possível dentro da disciplina, foram gerados diversos conjuntos de dados com base nas combinações incrementais das sete semanas de interações existentes para a disciplina. Assim, foram gerados conjuntos de dados contendo somente a primeira semana de interações, contendo a primeira e a segunda semana, e assim por diante, até ser gerado o conjunto de dados com todas as 7 semanas de interações. Um conjunto de dados consiste em todos os exemplos de alunos de um dado semestre de um curso.

3.3. Atributos Derivados

Ainda que a série temporal gerada pelo número absoluto de interações dos alunos possa ser útil, geramos também um conjunto de atributos derivados que podem auxiliar na contextualização desses valores. O conjunto total de atributos utilizados e a descrição dos mesmos são apresentados na Tabela 2. Todos atributos são derivados unicamente na contagem total de interações e relações entre interações de alunos, tutores e professores.

Este conjunto de atributos foi escolhido de modo que certos aspectos dos dados originais fossem enfatizados. Esse é o caso da média, mediana e número de semanas com zero interações. Os demais foram criados com base em hipóteses, tentando revelar características ainda não conhecidas ou relações entre os dados de professores e tutores

com os alunos. A principal hipótese é de que o número absoluto de interações dos estudantes tem pouco valor, sendo necessário colocá-lo em contexto do número de interações da turma como um todo e dos tutores e professores. Também foi buscada uma inter-relação entre os dados de um exemplo de aluno com os dados dos demais exemplos de aluno da disciplina, caso este do fator de empenho.

Tabela 2. Conjunto de Atributos

Atributo	**Descrição**
Situação Acadêmica	Situação final do aluno na disciplina (rótulo do exemplo)
Interações por Semana (1 até 7 semanas)	Número de interações na semana.
Média	Média do total de interações pelo número de semanas.
Mediana	Mediana do conjunto de interações por semana.
Semanas Zeradas	Número de semanas com zero interações.
Média da Diferença	Média da diferença entre a semana i e a semana $i+1$.
Razão com Professores	Razão entre o total de interações do aluno e dos professores.
Razão com Tutores	Razão entre o total de interações do aluno e dos tutores.
Fator de Empenho (1 até 7 semanas)	Razão entre as interações da semana do aluno e a média de interações da turma naquela semana.

4. Metodologia de Análise

Neste artigo, estamos interessados em verificar a eficácia de diferentes modelos em prever a situação de alunos em diferentes situações. Em particular, focamos em dois casos: a) utilizando somente o número absoluto de interações dos estudantes e b) adicionando os atributos derivados. O primeiro caso visa identificar o potencial do modelo de dados constituído somente pela situação do aluno e as interações que ele teve em cada semana. O segundo caso utiliza as mesmas informações do primeiro, porém acrescentando os atributos derivados (média, mediana, média da diferença, semanas zeradas, razão com os professores, razão com os tutores e o fator de empenho de cada semana).

Para cada análise, dividiu-se os conjuntos de exemplos em um conjunto de treinamento e um conjunto de teste. Esta divisão foi feita de duas formas para testar duas situações distintas. Na primeira forma, denominada "entre turmas", a divisão é feita dentro de um mesmo semestre do mesmo curso, utilizando turmas distintas para treino e teste. Na segunda forma, denominada "entre semestres", a divisão é feita entre semestres distintos.

Para tanto, cada conjunto de dados foi dividido aleatoriamente em dois conjuntos contendo a mesma quantidade de alunos, ou seja, cada turma de alunos foi transformada em duas turmas (A e B). No primeiro formato (entre turmas) uma turma (ex: turma A) de um semestre de um curso foi utilizada para treinar o modelo, e a outra turma (ex: turma B) daquele semestre daquele curso foi utilizada para avaliar o modelo. No segundo formato (entre semestres), cada uma das turmas (A e B) de cada semestre foi utilizada para treinar os modelos, e os mesmos foram avaliados com as turmas A e B do semestre seguinte (e vice-versa).

Esse processo de divisão foi repetido 5 vezes, gerando conjuntos aleatórios diferentes a cada repetição. As médias dos resultados da aplicação dos classificadores nestas repetições são reportadas.

Os seguintes modelos foram gerados e avaliados utilizando a ferramenta WEKA (Hall et al., 2009): Rede Bayesiana, Rede Neural (Perceptron de Múltiplas Camadas), J48 e Floresta Aleatória.

5. Resultados

De modo geral, as acurácias médias de classificação alcançaram altos índices já para os modelos que utilizavam somente os dados das interações da primeira semana, tendo sido possível, por exemplo, gerar modelos com acurácia média superior a 90%.

No entanto, medidas de acurácia dos classificadores não são adequadas aqui, já que as classes não são balanceadas - há muito mais candidatos retidos do que evadidos, fazendo com que classificadores que somente indicam retenção tenham já alta acurácia sem serem realmente úteis. Assim, focamos aqui na capacidade dos classificadores em indicar corretamente reprovações - isto é, a fração corretamente classificada de estudantes reprovados. A precisão na classificação de alunos aprovados manteve-se sempre alta, entre 75% e 95%, demonstrando que a tarefa de prever reprovação é substancialmente mais difícil.

Por restrições de espaço descreveremos somente os experimentos relacionados ao Curso de Licenciatura em Educação do Campo, mas resultados semelhantes foram obtidos para o curso de Pedagogia.

5.1 Classificação utilizando somente número de interações

Como é possível observar na Tabela 3, os modelos se tornam mais precisos conforme mais semanas são disponibilizadas ao modelo. Quando utilizamos somente as quantidades de interações da primeira semana, as precisões para classificar alunos reprovados são próximas a zero.

A medida que dados de mais semanas são adicionados, os modelos tornam-se, como esperado, mais precisos. Observa-se no entanto que há uma diminuição de ganhos com o passar das semanas - além da quinta semana novos dados acrescentam muito pouco ao modelo.

De forma geral, as Redes Bayesianas obtiveram os melhores resultados, com exceção da primeira e segunda semanas, onde as Florestas Aleatórias obtiveram melhores resultados.

Tabela 3. Precisões para classificação de reprovados – experimento sem utilização de atributos

Caso	Modelo	Semana						
		S1	S2	S3	S4	S5	S6	S7
Entre Semestres	Rede Bayesiana	0,0	0,22	0,44	0,51	0,60	0,63	**0,64**
	Rede Neural	0,0	0,16	0,33	0,45	0,49	0,51	0,51
	J48	0,0	0,21	0,37	0,40	0,52	0,56	0,56
	Floresta Aleatória	0,1	0,33	0,36	0,40	0,48	0,51	0,54
Entre Turmas	Rede Bayesiana	0,0	0,24	0,45	0,51	0,58	0,62	**0,66**
	Rede Neural	0,0	0,14	0,33	0,46	0,49	0,53	0,56
	J48	0,0	0,20	0,39	0,42	0,51	0,59	0,60
	Floresta Aleatória	0,1	0,28	0,33	0,41	0,50	0,56	0,61

5.2 Classificação utilizando atributos derivados

Para os modelos treinados com atributos derivados temos um comportamento semelhante aos do experimento anterior. De modo geral os modelos também tornam-se mais precisos a medida que as informações de interação de cada semana vão sendo disponibilizadas e Redes Bayesianas se mostraram o modelo mais adequado novamente. Observando apenas a última semana, não há ganhos na precisão da classificação.

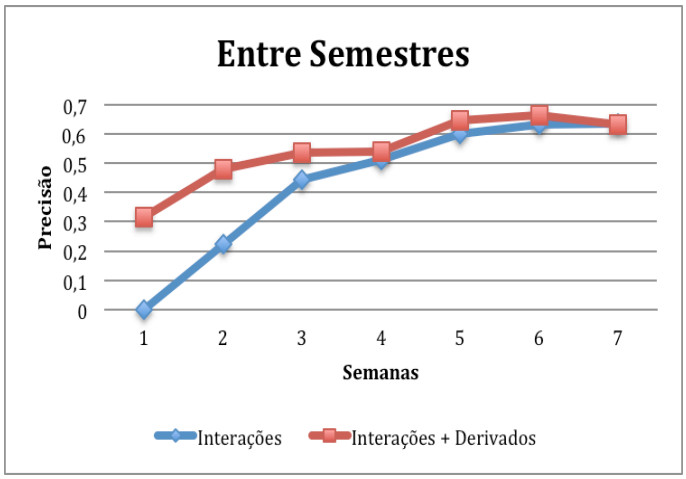
Tabela 4. Precisões para classificação de reprovados – experimento com a utilização de atributos

Testes	Modelo	Semana						
		S1	S2	S3	S4	S5	S6	S7
Entre Semestres	Rede Bayesiana	0,32	0,48	0,54	0,54	0,65	0,66	**0,63**
	Rede Neural	0,07	0,22	0,38	0,42	0,46	0,55	0,58
	J48	0,02	0,25	0,34	0,37	0,48	0,50	0,53
	Floresta Aleatória	0,06	0,26	0,30	0,37	0,41	0,47	0,46
Entre Turmas	Rede Bayesiana	0,35	0,48	0,52	0,57	0,64	0,68	**0,67**
	Rede Neural	0,08	0,24	0,41	0,44	0,49	0,57	0,59
	J48	0,05	0,27	0,36	0,39	0,49	0,57	0,64
	Floresta Aleatória	0,11	0,27	0,33	0,41	0,45	0,54	0,51

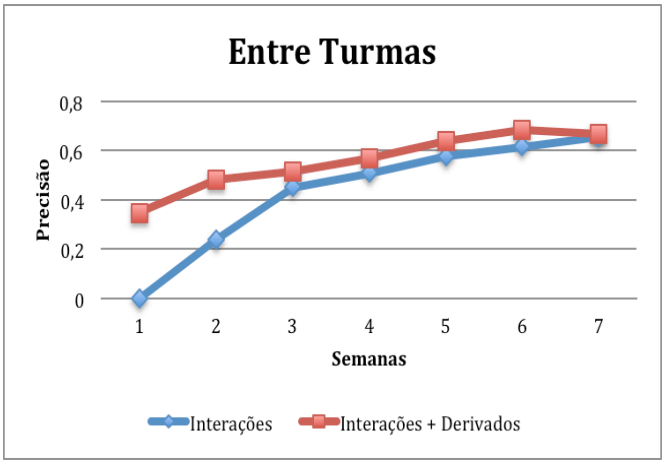
5.3 Comparação dos resultados

A Figura 1 compara o desempenho obtido por semana utilizando Redes Bayesianas e as duas combinações de atributos. Observa-se que, para as semanas finais, as precisões dos modelos são virtualmente idênticas. No entanto, para semanas iniciais, a utilização dos atributos derivados torna-se substancialmente vantajosa. A partir da quarta semana os resultados se aproximam, convergindo para essencialmente a mesma precisão na última semana.

Figura 1. Comparação de desempenho a cada semana utilizando Redes Bayesianas e diferentes combinações de atributos



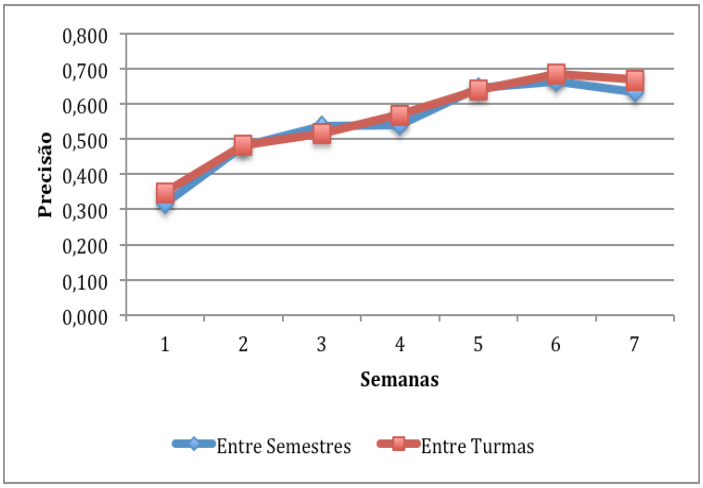
(a) Comparação entre semestres



(b) Comparação entre turmas

Observa-se ainda que a precisão a cada semana é essencialmente a mesma independente se o modelo foi treinado entre turmas ou entre semestres (Figura 2). Este é um resultado interessante, pois demonstra uma invariância no resultado em relação a origem do treinamento. Isto é, um modelo pode ser treinado com dados de um semestre para ser aplicado no semestre posterior, bem como treinado em um conjunto de estudantes para ser aplicado em outros estudantes.

Figura 2. Comparação de precisão por semana utilizando Rede Bayesiana e todos atributos disponíveis



6. Considerações Finais e Trabalhos Futuros

A possibilidade de prever com antecedência se um estudante de educação a distância corre o risco de não concluir uma disciplina ou curso é de grande valia para professores e tutores, que podem ajustar seus instrumentos pedagógicos para evitar que estes estudantes reprovem ou evadam.

Neste artigo mostramos resultados da aplicação de técnicas de aprendizado de máquina na predição de reprovação de estudantes. Demonstramos que utilizar apenas a quantidade de interações dos alunos é viável para gerar predições razoavelmente precisas, ainda que menos precisas do que trabalhos anteriores que utilizam atributos muito mais específicos; estes, porém, sacrificam generalidade na aplicação em outros contextos, enquanto nossa abordagem é aplicável virtualmente em qualquer situação onde é possível contar interações de qualquer tipo.

Mostramos ainda que utilizar apenas o número de interações dos estudantes leva a uma baixa capacidade preditiva quando poucas semanas de dados estão disponíveis. Derivando-se atributos que levam em conta a quantidade de interações de professores e tutores, bem como outros atributos comparativos, e utilizando estes atributos nos modelos, obtém-se um desempenho bastante superior para semanas iniciais ao utilizar-se um modelo capaz de aproveitar estar informações. Observamos que estas derivações não violam nosso requisito de apenas utilizar contagens de interações genéricas, já que ainda não especificamos em nenhum momento que interações estão sendo realizadas.

Por fim, mostramos que redes bayesianas se mostraram o modelo mais adequado para o problema e que é possível treinar o modelo utilizando diferentes fontes de exemplos. Em particular, obtivemos precisões idênticas ao treinar tanto entre turmas, onde o treino é realizado sobre um conjunto de estudantes e a aplicação em outro conjunto disjunto, como entre semestres, onde o treino ocorre utilizando estudantes de um semestre e aplica-se em outro semestre.

Trabalhos futuros incluem reduzir a diferença de desempenho quando utilizando-se apenas contagem de interações e utilizando-se atributos mais específicos. Para isso, parece ser necessário derivar outros atributos ou encontrar melhores modelos capazes de

fazer uso destes atributos. Adicionalmente, é necessário ampliar os testes aqui realizados para outros cursos, inclusive em outras instituições, de forma a melhor validar o modelo.

Referências

Brown, M. (2012). Learning Analytics: Moving from Concept to Practice. *EDUCAUSE Learning Initiative Brief*.

CensoEaD (2013), from http://www.abed.org.br/censoead/censoEAD.BR_2012_pt.pdf.

Gottardo, E., Kaester, C., & Noronha, R. V. (2012). Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. *Anais do XXIII SBIE*.

Gotardo, R., Cereda, P., Hruschka Junior, E. (2013). Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores. Anais do XXIV Simpósio Brasileiro de Informática na Educação.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten. I.H. (2009) "The WEKA Data Mining Software: An Update" SIGKDD Explorations, Volume 11, Issue 1.

Hämäläinen, W., Vinni, M. (2011) "Classifiers for Educational Data Mining". In: Romero et al. Handbook of Educational Data Mining. Flórida, CRC Press, p. 57-71.

Macfadyen, L.P., Dawson, S. (2010) "Mining LMS Data to Develop an "Early Warning System" for Educators: A Proof of Concept". Computers & Education, no. 54, p.588-599.

Manhaes, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrao, G. (2011). Previsão de Estudantes com Risco de Evasãoo Utilizando Técnicas de Mineração de Dados. *Anais do XXII SBIE*.

Mishra, T.; Kumar, D.; Gupta, S., "Mining Students' Data for Prediction Performance," Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on , vol., no., pp.255,262, 8-9 Feb. 2014

Rodrigues, R. L., de Medeiros, F. P. A., & Gomes, A. S. (2013). Modelo de Regressão Linear aplicador à previsão de desempenho de estudantes em ambiente de aprendizagem. *Anais do XXIV SBIE*.

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40*(6), 601-618.