Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina

Daniel Miranda de Brito[1], Iron Araújo de Almeida Júnior[1], Eduardo Vieira Queiroga[1], Thaís Gaudencio do Rêgo[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB) – João Pessoa – PB – Brasil

{britmb,ironaraujo,gaudenciothais}@gmail.com, eduardo.queiroga@ci.ufpb.br

Abstract. *College students drop-out presents a sobering statistic faced by universities. The problem of failure of students in early periods is often considered as an influential factor on drop-out. This paper proposes the use of Data Mining techniques to try to predict the performance of Computer Science students in the first semester, at UFPB, through their notes of entrance exam. The results showed that it is possible to infer the performance of students with a precision better than 70%, with this information useful for performing actions to prevent drop-out, improving the education system.*

Resumo. *A evasão de estudantes universitários apresenta uma estatística preocupante enfrentada pelas universidades. O problema da reprovação dos estudantes em períodos iniciais é muitas vezes considerado como um fator influenciador da evasão. Este trabalho propõe a utilização de técnicas de Mineração de Dados para tentar predizer o desempenho dos alunos no primeiro período do curso de Ciência da Computação da UFPB, através das suas notas de ingresso no vestibular. Os resultados mostraram que é possível inferir o desempenho dos estudantes com uma acurácia superior a 70%, sendo esta informação útil para a realização de ações para evitar a evasão, aprimorando o sistema de ensino.*

1. Introdução

O acesso da população ao ensino superior no Brasil tem crescido nos últimos anos, porém, este crescimento também traz consigo uma estatística preocupante: o alto índice de evasão de alunos [Reis et al. 2012]. A busca de causas para esse problema tem sido objeto de estudo de muitos trabalhos e pesquisas, como pode ser visto em Silva Filho et al. (2007). Barroso e Falcão (2004) discutem os tipos de causas de evasão em um estudo realizado no curso de Física da UFRJ: evasão econômica, onde existe a impossibilidade de vínculo devido a questões socio-econômicas; evasão vocacional, onde o aluno abandona o curso por falta de identificação com o mesmo; evasão institucional, onde o aluno abandona o curso por fracasso nas disciplinas iniciais, seja por deficiências prévias de conteúdo, inadequação nos métodos de estudo, ou por dificuldade de relacionamento com outros colegas. Este último tipo de evasão deve ser combatido com ações pedagógicas por parte das instituições. Ainda em Silva Filho et al. (2007), discute-

se que a taxa de evasão em todo o mundo é maior de duas a três vezes no primeiro ano do curso.

As ações realizadas pelas instituições para ajudar estudantes do primeiro ano são geralmente atividades de tutoria e monitoria [Veenstra 2009], porém, Seidman (2005) sugere que haja uma identificação precoce e assistência intensiva aos estudantes que correm risco de retenção ou evasão. Nos últimos anos, técnicas de Mineração de Dados (do inglês *Data Mining*, DM) têm sido utilizadas como forma de predição de desempenho dos estudantes. DM consiste na aplicação de algoritmos específicos para extrair padrões a partir de dados. A DM é parte de um contexto mais amplo, conhecido como Descoberta de Conhecimento em Banco de Dados (do inglês *Knowledge Discovery in Databases*, KDD). KDD é o processo de descobrir conhecimento útil a partir de dados e além da etapa de DM, envolve etapas adicionais, como preparação, seleção e limpeza dos dados, incorporação de conhecimento prévio apropriado e interpretação adequada dos resultados da mineração de dados, a fim de garantir que resultados úteis sejam derivados dos dados [Fayyad et al. 1996]. O conhecimento obtido a partir de dados pode ser útil para o aprimoramento dos sistemas de ensino, originando uma área de pesquisa denominada Mineração de Dados Educacionais (do inglês *Educational Data Mining*, EDM). EDM está preocupada com o desenvolvimento de métodos para explorar informações coletadas de ambientes educacionais, permitindo compreender os alunos de forma mais eficaz e adequada, proporcionando melhores benefícios educacionais aos mesmos [Baker et al. 2011].

O objetivo deste trabalho é identificar os estudantes que necessitam de apoio didático nas disciplinas da área de exatas (Cálculo Diferencial e Integral I, Física Aplicada à Computação I, Cálculo Vetorial e Geometria Analítica) do primeiro período do curso de Ciência da Computação da Universidade Federal da Paraíba (UFPB). Através de um conjunto de dados real, foi avaliada a relação entre as notas de ingresso do aluno e o seu desempenho no primeiro período do curso. Os resultados deram indícios de que essa relação existe, sendo uma informação relevante para que medidas sejam tomadas para a diminuição da retenção ou evasão acadêmica ao suprir deficiências provenientes do ensino médio.

O artigo está organizado da seguinte maneira: Na Seção 2 apresentam-se os trabalhos relacionados; na Seção 3 descrevem-se como os dados foram obtidos e o pré-processamento realizado nos mesmos; os resultados são apresentados na Seção 4 e as considerações finais e trabalhos futuros na Seção 5.

2. Trabalhos Relacionados

Vários autores realizam análises e propõem ferramentas no contexto de DM na educação. Romero et al. (2008) comparam técnicas de DM na classificação de estudantes com base nos dados de uso e nas notas finais do curso na plataforma *moodle* de gerenciamento de aprendizagem. Também foi desenvolvida uma ferramenta de apoio a tomada de decisão para uso dos instrutores baseada nas técnicas analisadas. Baradwaj e Pal (2011) propuseram um modelo de DM para analisar a performance de alunos de ensino superior, utilizando os dados de desempenho do fim do semestre em uma tarefa de classificação com árvore de decisão. Talaveria et al. (2004) buscaram encontrar grupos de comportamentos distintos entre estudantes em ambientes virtuais de

aprendizado. Veenstra (2009) propôs um *framework* para direcionar ações de apoio para os alunos calouros, que necessitam de suporte de acordo com suas características acadêmicas do ensino médio.

Dekker et al. (2009) compararam diversos algoritmos classificadores da ferramenta de mineração de dados Weka, na tentativa de prever o risco de evasão de estudantes com base no desempenho do primeiro semestre letivo. Foram utilizados dados de estudantes do curso de Engenharia Elétrica da Universidade de Tecnologia de Eindhoven. O classificador árvore de decisão apresentou os melhores resultados, com uma acurácia entre 75 e 80%.

No Brasil, as oportunidades para a prática da EDM vêm crescendo muito. A criação dos cursos de Educação à distância (EaD) criou várias oportunidades para as pesquisas na área [Baker et al. 2011]. Sales et al. (2011) identificaram variáveis preditivas relacionadas a persistência e evasão em treinamentos a distância, enquanto Gottardo et al. (2012) estudaram e comprovaram a viabilidade da realização de inferências relativas ao desempenho de estudantes em cursos EaD. Kampff et al. (2008) buscou identificar alunos com risco de evasão ou reprovação através de técnicas de DM, auxiliando os professores a atuarem de forma a evitar esses resultados. No trabalho de Manhães (2011), as notas dos alunos calouros do curso de Engenharia Civil da UFRJ são utilizadas para verificar a sua situação final no curso, identificando se o mesmo possui risco de evasão. Neste estudo, os autores obtiveram uma acurácia média variando de 75% a 80%. Já de França e do Amaral (2013) fizeram uso de técnicas de DM para identificar grupos similares de estudantes com dificuldade na disciplina de programação, com o intuito de que estratégias pedagógicas fossem planejadas para os grupos de estudantes que apresentassem dificuldades.

3. Obtenção dos dados e Pré-processamento

Para a realização deste estudo utilizou-se um conjunto de dados real de alunos do curso de Ciência da Computação da UFPB, entre o ano de 2006 e 2013. Esses dados foram fornecidos pela Superintendência de Tecnologia da Informação (STI) da UFPB e consistem nas notas de ingresso e de disciplinas pertencentes ao primeiro período do curso. As disciplinas consideradas na análise são: Cálculo Diferencial e Integral I, Física Aplicada à Computação I, Cálculo Vetorial e Geometria Analítica. Estas disciplinas pertencem a um grupo que historicamente possui uma elevada taxa de reprovação em cursos da área de Exatas [Barbosa 1995; Fernandes Filho 2001; Soares 2011].

As médias finais dos alunos nestas disciplinas foram utilizadas como critério para a seleção das classes de desempenho, conforme discutido mais adiante. Como o fracasso nas disciplinas iniciais do curso pode estar relacionado com deficiências prévias de conteúdo, optou-se por selecionar as informações relacionadas ao desempenho do vestibular do aluno essenciais para as disciplinas utilizadas como parâmetro de desempenho. A média geral obtida no vestibular também foi selecionada como forma de avaliação do conhecimento geral do aluno nas diversas áreas. Os atributos de ingresso utilizados foram: Média Geral, Média de Matemática e a Média de Física obtidas no processo seletivo para entrada na UFPB.

Uma análise da correlação entre os atributos de ingresso foi realizada para identificar a existência de possíveis redundâncias entre os mesmos, proveniente de uma

correlação positiva muito alta. Na Figura 1(a), apresenta-se o gráfico de dispersão dos atributos Média de Matemática e a Média de Física no vestibular com um coeficiente de correlação r = 0,62, o que indica uma correlação positiva moderada. Na Figura 1(b), para os atributos Média Geral e Média de Matemática obtidas no vestibular, tem-se r = 0,74, o que indica uma correlação positiva alta. A Figura 1(c), para os atributos Média Geral e Média de Física no vestibular temos r = 0,76, que representa uma correlação positiva alta. Nenhum dos possíveis pares de atributos possui uma correlação positiva muito alta (superior a 0,9), indicando a relevância individual dos mesmos na tarefa de classificação dos dados. Além disso, testes foram realizados com a exclusão de cada uma das variáveis, onde foi observada a relevância dos três atributos na tarefa de classificação.

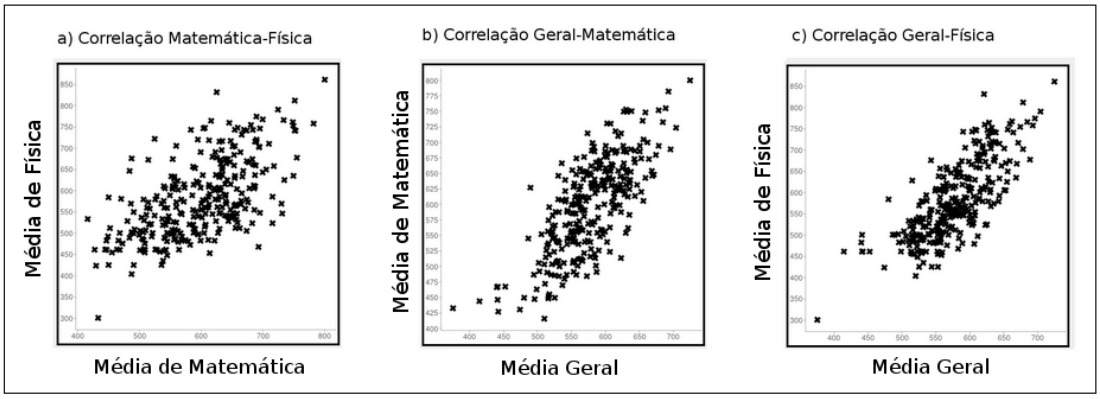

Figura 1. Gráficos de Dispersão para correlação entre os atributos de ingresso

Alunos que apresentaram ausência de alguma das notas, ou seja, alunos que trancaram a disciplina, ainda estão cursando, ou devido a algum problema na base de dados inexiste nela, foram desconsiderados na etapa de pré-processamento, resultando em um conjunto de 300 instâncias. Os estudantes foram divididos em duas classes distintas, de acordo com a quantidade de aprovações nas disciplinas cursadas. Caso o estudante tenha sido aprovado em todas as três disciplinas, isto é, possui as médias maiores ou iguais a 5,0, é considerado como sendo da classe "A", e caso tenha reprovado (nota inferior a 5,0 ou por falta) em pelo menos uma das três disciplinas, é considerado como sendo da classe "B". A Tabela 1 mostra a divisão dos estudantes de acordo com o tipo de classe.

Tabela 1. Divisão dos estudantes em classes

Classe	Descrição	Número de Instâncias
A	Estudantes aprovados em todas as disciplinas	138
B	Estudantes que reprovaram pelo menos uma disciplina	162

4. Discussão dos resultados

Para este trabalho, foram utilizados algoritmos de aprendizado de máquina, dentre os implementados pela ferramenta WEKA (*Waikato Environment for Knowledge Analysis*)

na versão 3.6.10, que permite que usuários experimentem e comparem resultados dos diferentes métodos [Hall et al. 2009].

Para a realização dos testes foi utilizada a interface gráfica *Explorer* do Weka. Esta interface permite a exploração dos dados e suporta o recurso de classificação [Bouckaert et al. 2010]. Foram utilizados cinco algoritmos de aprendizado de máquina pertencentes a classes distintas de classificadores, onde para cada um deles foram realizadas três variações de parâmetros, com exceção do algoritmo *NaiveBayes*, que não permite possibilidades de variação, conforme pode ser visto na Tabela 2.

Tabela 2. Algoritmos de aprendizado de máquina utilizados

Algoritmo	Classe
Naive Bayes	Métodos Bayesianos
IBk	Métodos de Vizinho mais Próximo
SMO	Máquina de Vetor de Suporte
Random Forest	Árvore de Decisão
Multilayer Perceptron	Redes Neurais Artificiais

O algoritmo classificador IBk foi executado com o parâmetro *k* (quantidade de vizinhos analisados) com valores 9, 11 e 13. O representante da Máquina de Vetor de Suporte teve a sua constante e o parâmetro que indica tolerância variados em (3,0 e 0,002), (1,0 e 0,001) e (5,0 e 0,001), respectivamente. A Árvore de Decisão teve o parâmetro que indica a quantidade de árvores usadas no classificador variado em 8, 10 e 15. E por fim, no algoritmo classificador representante das redes neurais alterou-se os parâmetros que indicam a taxa de aprendizado e o momento entre (0,3; 0,2), (0,4; 0,3) e (0,5; 0,4), respectivamente.

A divisão entre o conjunto de treinamento e testes foi feita utilizando a Validação Cruzada (do inglês, *Cross Validation*). Nesta abordagem, o conjunto de dados é dividido em 10 partes iguais (*Folds*), sendo 9 destas utilizadas para treinamento e a restante utilizada para teste, a fim de se criar um modelo. O processo é repetido 10 vezes no total, usando um segmento diferente de cada vez para o teste, totalizando 10 modelos de onde extrai-se a média representativa de todos os resultados dos modelos. Como forma de avaliação do desempenho dos algoritmos, utilizou-se a precisão e verdadeiros positivos das duas classes. Os dados das execuções podem ser vistos na Tabela 3.

Observando os resultados obtidos mostrados na Tabela 3, verifica-se que a acurácia dos algoritmos nunca foi menor do que 70%. O algoritmo que obteve maior precisão foi o SMO, com média de 74,67%. Pode-se observar ainda que os alunos da classe A (aprovados nas 3 disciplinas) são mais difíceis de serem classificados do que os alunos da classe B (reprovados em pelo menos 1 disciplina). Acredita-se que isto ocorra pelo fato do bom desempenho do aluno no vestibular não ser o único fator influenciador no sucesso do aluno no primeiro período do curso, apesar de ser uma característica extremamente importante no seu sucesso.

Tabela 3. Precisão dos algoritmos de aprendizagem de máquina

	NaiveBayes	IBk	SMO	Random Forest	Multilayer Perceptron
Precisão	75%	73,33%	74,67%	71,67%	73,55%
Desvio Padrão	--	0,72%	0,72%	1,63%	0,41%
Verdadeiros Positivos classe "A"	73,2%	68,37%	66,9%	67,37%	71%
Desvio Padrão	--	1,23%	1,79%	2,58%	0,57%
Verdadeiros Positivos classe "B"	76,5%	77,57%	81,3%	75,3%	75,7%
Desvio Padrão	--	1,28%	0,28%	1,85%	0,28%

O gráfico de barras contendo a média da precisão dos algoritmos é mostrado na Figura 2.

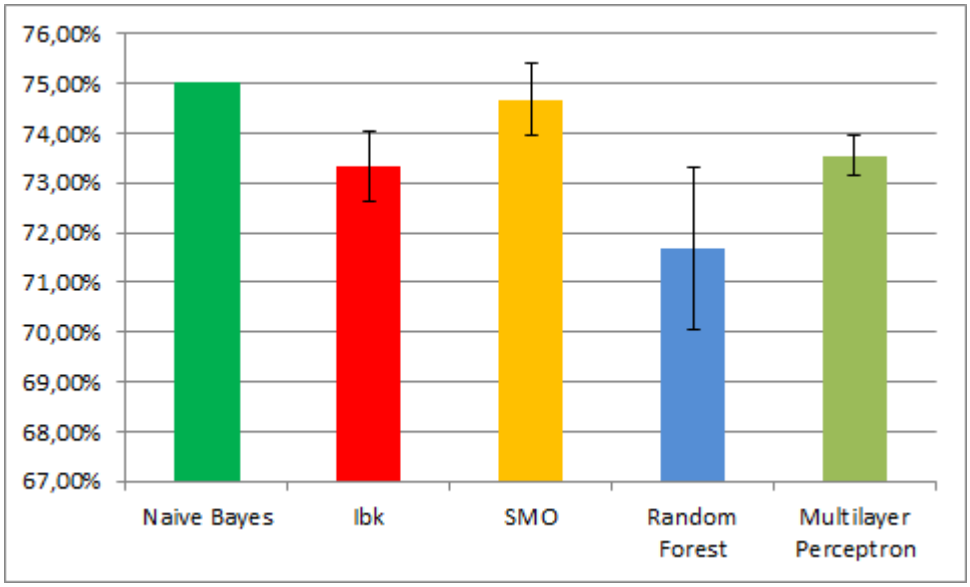

Figura 2. Taxas de acerto dos algoritmos

A taxa de verdadeiros positivos (que representa a proporção de indivíduos de uma classe que foi classificada de forma correta) para a classe B é maior do que a taxa de verdadeiros positivos para a classe A. O fato da maioria dos alunos da classe B terem sido classificados corretamente é relevante, dado que estes representam o conjunto de alunos que precisam de um maior apoio didático e consequentemente, apresentam um maior risco de evasão por fracasso nas disciplinas do primeiro período. Como contribuição para o ensino universitário, em relação a outros trabalhos da área, citados

na seção de trabalhos relacionados, tem-se que utilizando esta abordagem é possível direcionar ações pedagógicas para alunos de risco de evasão ou retenção, antes mesmo do início do primeiro período do curso, apenas com as notas de entrada do vestibular. Nos trabalhos relacionados, as previsões de desempenho dos alunos são feitas a partir de dados do desempenho desde o primeiro período do curso.

5. Considerações Finais

Este estudo buscou encontrar relação entre a nota de ingresso de estudantes e o seu desempenho nas disciplinas de Cálculo Diferencial e Integral I, Física Aplicada à Computação I, Cálculo Vetorial e Geometria Analítica do primeiro período do curso de Ciência da Computação da UFPB. Através da ferramenta Weka, obteve-se precisão superior a 70%, utilizando um conjunto de três atributos de entrada: Média Geral, Média de Matemática e a Média de Física obtidas no processo seletivo para entrada na UFPB. Estes resultados servem de indício para mostrar que é viável realizar a predição do desempenho baseado em suas notas de ingresso, o que permite aos educadores tomarem providências com o objetivo de contornar o baixo rendimento estudantil. Tem-se conhecimento da existência de outras variáveis que podem influenciar o desempenho do aluno no primeiro período do curso, porém estas são muitas vezes subjetivas e difíceis de serem recuperadas, como motivação do aluno no curso, taxa de aprovação da turma, situação socioeconômica, entre outras.

Acredita-se que os resultados obtidos neste estudo possam ajudar os educadores, uma vez que é possível obter estimativas sobre o desempenho dos alunos, e então servir de base para o planejamento de estratégias e políticas que visem diminuir o número de reprovações, reduzindo, como consequência, a evasão dos alunos do curso de Ciência da Computação. Como trabalhos futuros, considera-se que uma análise mais detalhada no processo de seleção dos atributos de entrada é essencial, investigando a existência de outras variáveis que possam influenciar no desempenho do aluno no primeiro período do curso. Estudar a relação entre o desempenho dos alunos no primeiro período e o desempenho destes em períodos subsequentes, também se faz necessário. Além disso, acredita-se que é possível realizar pesquisa semelhante para alunos de cursos diferentes, verificando se os resultados são semelhantes para outras disciplinas presentes no vestibular.

Referências

Baker, R., Isotani, S. e Carvalho, A. (2011) “Mineração de Dados Educacionais: Oportunidades para o Brasil”. Revista Brasileira de Informática na Educação, v. 19, n. 02, pp. 3-13.

Baradwaj, B. K. e Pal, S. (2011) “Mining educational data to analyze students’ performance”. International Journal of Advance Computer Science and Applications, v. 2, n. 6, pp. 63-69.

Barbosa, G. O. e Hermínio, B. N. (1995) “Raciocínio lógico formal e aprendizagem em cálculo diferencial e integral: o caso da Universidade Federal do Ceará”. Tema e Debates, n. 6, pp. 60-70.

Barroso, M. F. e Falcão, E. B. (2004) “Evasão universitária: o caso do Instituto de Física da UFRJ”. IX Encontro Nacional de Pesquisa em Ensino de Física.

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P. e Witten, I. H. (2010) "WEKA---Experiences with a Java Open-Source Project". The Journal of Machine Learning Research, v. 11, pp. 2533-2541.

de França, R. S. e do Amaral, H. J. C. (2013) "Mineração de Dados na Identificação de Grupos de Estudantes com Dificuldade de Aprendizagem no Ensino da Programação". RENOTE, v. 11, n. 1.

Dekker, G. W., Pechenizky, M. e Vleeshouwers, J. W. (2009) "Predicting Students Drop Out: A Case Study". International Working Group on Educational Data Mining.

Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. (1996) "From data mining to knowledge discovery in databases". Al magazine, v. 17, n. 3, pp. 37-54.

Fernandes Filho, O. P. (2001) "O desenvolvimento cognitivo e a reprovação no curso de engenharia". In XXIX Congresso Brasileiro de Ensino da Engenharia.

Gottardo, E., Kaestner, C. e Noronha, R. V. (2012) "Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais". In Anais do Simpósio Brasileiro de Informática na Educação, v. 23, n. 1.

Hall, M., Frank, E, Holmes, G, Pfahringer, B., Reutemann, P. e Witten, I. H. (2009) "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter, v. 11, n. 1, pp. 10-19.

Kampff, A. J. C., Reategui, E. B. e de Lima, J. V. (2008) "Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente". RENOTE, v. 6, n. 1.

Manhães, L. M. B., da Cruz, S. M. S., Macário Costa, R. J., Zavaleta, J. e Zimbrão, G. (2011) "Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados". In Anais do Simpósio Brasileiro de Informática na Educação, v. 22, n. 1, pp. 150-159.

Reis, V. W., Cunha, P. J. e Spritzer, I. M. (2012) "Evasão no ensino superior de engenharia no Brasil: um estudo de caso no CEFET/RJ". Belém: Cobenge. Disponível em: <http://www.abenge.org.br/CobengeAnteriores/2012/artigos/103734.pdf>. Acesso em: 15 jun. 2014.

Romero, C., Ventura, S., Espejo, P. G. e Hervás, C (2008) "Data Mining Algorithms to Classify Students". In EDM, pp. 8-17.

Romero, C. e Ventura, S. (2010) "Educational data mining: a review of the state of art". Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, v. 40, n. 6, pp. 601-618.

Sales, P. A. O., Abbad, G. e Rodrigues, J. L. (2011) "Variáveis Preditivas de Evasão e Persistência em Treinamentos a Distância". XXV Encontro da ANPAD.

Seidman, A. (2005). "College student retention: Formula for student success". Greenwood Publishing Group.

Silva Filho, R. L. L., Montejunas, P. R., Hipólito, O. e Lobo, M. B. C. M. (2007) "A evasão no ensino superior brasileiro". Cadernos de Pesquisa, v. 37, n. 132, pp. 641-659.

Soares, F. S. (2011) "Levantamento e Análise de Dados Sobre a Evasão no Curso de Licenciatura em Computação do Campus VII da Universidade Estadual da Paraíba – UEPB". Trabalho de Conclusão de Curso.

Talavera, L. e Gaudioso, E. (2004) "Mining student data to characterize similar behavior groups in unstructured collaboration spaces". In Workshop on artificial intelligence in CSCL, 16th European conference on artificial intelligence, pp. 17-23.

Veenstra, C. P. (2009) "A strategy for improving freshman college retention". Jornal for Quality and Participation, v. 31, n. 4, pp. 19-23.