Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de Data Mining

Daniel Miranda de Brito
Programa de Pós-Graduação em Informática
Universidade Federal da Paraíba (UFPB) – João Pessoa – PB, Brasil
britmb@gmail.com

Túlio Albuquerque Pascoal
Centro de Informática
Universidade Federal da Paraíba (UFPB) – João Pessoa – PB, Brasil
tuliopascoal@gmail.com

Jairo Gustavo G. de O. Araújo
Centro de Informática
Universidade Federal da Paraíba (UFPB) – João Pessoa – PB, Brasil
jairogustavo007@gmail.com

Marcílio O. Lemos
Programa de Pós-Graduação em Informática
Universidade Federal da Paraíba (UFPB) – João Pessoa – PB, Brasil
marcilio.cc.lemos@gmail.com

Thaís Gaudencio do Rêgo
Centro de Informática
Universidade Federal da Paraíba (UFPB) – João Pessoa – PB, Brasil
gaudenciothais@gmail.com

ABSTRACT
Educators have been concerned about the high dropout rates in higher education in Brazil, responsible for social and economic losses to the country. This problem is more evident in Computer courses where dropout rates are even higher. In this paper, we propose a method for identifying students with risk of dropout from the analysis of the admission test scores and the grades of the first semester of the course. The results show that the potential dropout students can be identified with success rates of up to 86.9%. Thus, we believe that this is a viable approach to help teachers actions aimed at the reduction of dropout rates.

RESUMO
Educadores têm se preocupado com os altos índices de evasão nos cursos superiores no Brasil, responsáveis por perdas sociais e econômicas ao país. Esse problema é mais evidente nos cursos de Computação, onde os índices de evasão são ainda maiores. Neste trabalho, propõe-se um método para a identificação de estudantes com risco de evasão a partir da análise das notas obtidas pelos estudantes na prova de ingresso na instituição e nas disciplinas do primeiro semestre do curso. Os resultados obtidos mostram que os potenciais alunos evadidos podem ser identificados com taxas de acerto de até 86,9%. Dessa maneira, acredita-se que esta é uma abordagem viável para auxiliar ações docentes que visem a diminuição dos índices de evasão.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data Mining; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems; K.3.2 [**Computers and Education**]: Computer science education.

General Terms
Design, Experimentation, Human Factors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
TISE'15, December 1–3, 2015, Santiago, Chile
Copyright 2015 ACM 1-58113-000-0/00/0010 …$15.00.

Keywords
Educational Data Mining, Machine Learning, Prediction, College dropout, Dropout prevention.

1. INTRODUÇÃO
Os altos índices de evasão nos cursos superiores do Brasil sejam nos sistemas de ensino públicos ou privados, têm preocupado educadores. Este problema é responsável por importantes perdas sociais, acadêmicas e econômicas. No entanto, a evasão não é um problema que se restringe ao Brasil e afeta diversas instituições em todo o mundo. A busca de causas e soluções para o problema é objeto de estudo de muitos trabalhos e pesquisas em geral [1], [20].

Segundo Silva Filho et al. [20], instituições de ensino superior e seus alunos dão como principal justificativa para a evasão a falta de recursos financeiros, que impedem a continuidade do curso, porém esta não é a única causa. Os principais fatores, na maior parte das vezes, estão relacionados à questões acadêmicas, expectativas do aluno em relação a formação e integração do estudante com a instituição. Barroso e Falcão [4] discutem em seu trabalho realizado no curso de Física da Universidade Federal do Rio de Janeiro (UFRJ), que a evasão destes alunos também pode estar associada ao fracasso nas disciplinas iniciais, seja por dificuldades prévias de conteúdo ou por dificuldades de relacionamento com os colegas e instituição. As autoras discutem que este tipo de evasão pode ser evitado por uma atuação docente bem planejada.

Em todo o mundo, a taxa de evasão no primeiro ano do curso é de duas a três vezes maior do que nos anos seguintes [20], logo, é importante o desenvolvimento de ferramentas para identificar precocemente os estudantes com risco de evasão, evitando-a, antes que ocorra de fato. Usando dados passados, é possível traçar o perfil de estudantes antigos, que não foram bem sucedidos na Universidade, permitindo que estudantes com perfis semelhantes sejam identificados e acompanhados prematuramente [19]. O uso de métodos computacionais pode auxiliar a identificação destes estudantes.

A Mineração de Dados Educacionais (do Inglês, *Educational Data Mining,* EDM) é uma área de pesquisa que tem como foco o

desenvolvimento de métodos para explorar conjuntos de dados coletados de ambientes educacionais [3]. A utilização destes métodos possibilita o melhor entendimento dos estudantes e do cenário em que eles aprendem [2]. A EDM consiste na aplicação de técnicas de Mineração de Dados (do Inglês, *Data Mining*, DM) no âmbito educacional. A DM é a aplicação de algoritmos específicos para extração de padrões de conjunto de dados e é uma etapa do processo global de Descoberta de Conhecimento em Banco de Dados (do Inglês, *Knowledge Discovery in Databases*, KDD), cujo objetivo é a descoberta de conhecimento útil em dados [10].

A Predição é uma das técnicas mais populares de EDM e tem como objetivo o desenvolvimento de modelos que permitam inferir um único aspecto dos dados (variável preditiva), a partir da combinação de outros aspectos (variáveis preditoras) [2]. Um modelo de classificação (classificador) é um mapeamento de instâncias para classes [9]. Alguns modelos produzem como saída valores contínuos (numéricos) e outros produzem valores discretos (rótulos de classe). Em EDM, os aspectos inferidos são normalmente desempenho e notas dos alunos [16]. Diversos métodos de aprendizagem de máquina foram aplicados na predição do desempenho de estudantes, que consiste em umas das mais antigas aplicações de DM na educação.

As áreas de Ciências, Matemática e Computação apresentam taxas de evasão consideravelmente acima da média de outros cursos [20]. Em razão disto é fundamental que docentes dos cursos destas áreas tenham auxílio de métodos e ferramentas para identificação precoce dos estudantes com risco de evasão, de forma a facilitar o direcionamento de ações para evitar o problema. Apresenta-se neste trabalho uma estratégia de apoio aos docentes na detecção precoce de estudantes de computação, com risco de evasão, baseada nas suas notas obtidas nas matérias do vestibular e nas disciplinas do primeiro semestre do curso. Para estudar a viabilidade do modelo na predição da situação final dos estudantes, utilizam-se dados do histórico de alunos do curso de Ciência da Computação, da Universidade Federal da Paraíba (UFPB). Este trabalho está organizado da seguinte forma: Na seção 2 apresentam-se os trabalhos relacionados; Na seção 3 discute-se a metodologia aplicada; Na seção 4, os resultados obtidos são mostrados e, por fim, na seção 5, expõem-se as conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

A alta disponibilidade de informações relacionadas ao desempenho de estudantes em cursos à distância (EaD) cria um campo profícuo para a aplicação de técnicas de EDM [11]. Vários autores se empenham na busca de alternativas para a previsão de desempenho de estudantes, identificando possíveis propensos a evasão e contribuindo com os sistemas de ensino no combate do problema. Kotsiantis et al. [14] utilizaram técnicas de aprendizagem de máquina para predição de estudantes com alta probabilidade de evasão em um curso de informática à distância. Os autores avaliaram a acurácia da predição em dois momentos. Utilizando dados sociais dos alunos e notas de avaliações até o meio do período, os autores obtiveram precisão de até 63% e 83%, respectivamente. Romero et al. [17] realizaram experimentos, com dados de alunos de vários cursos à distância, para avaliar a eficiência e utilidade dos algoritmos de classificação na predição das notas finais, baseado em estatísticas de uso do sistema pelos alunos. Nos testes realizados, os melhores algoritmos resultaram em taxas de acerto próximas a 60%.

A popularização da modalidade de ensino à distância no Brasil favoreceu à pratica de DM em ambientes educacionais [3]. Gottardo et al. [12] investigaram a possibilidade de aplicação de técnicas de Mineração de Dados para previsão de desempenho de alunos no *Moodle*, a partir de diversos dados de acesso ao sistema. Através do uso de vários algoritmos de aprendizagem de máquina, os autores concluíram que é possível prever o desempenho dos mesmos com precisão superior a 74%. Santos et al. [18] aplicaram técnicas de Mineração de Dados para identificar estudantes com risco de reprovação na disciplina de Introdução à Arquitetura de Computadores, a partir o resultado das avaliações somativas, realizadas no decorrer do curso. As classificações corretas chegaram a 69%.

Diversos trabalhos de EDM também foram aplicados em ambientes educacionais presenciais. Dekker et al. [8] consideraram em seu estudo dados da educação pré-universitária e notas dos primeiros anos da universidade, de alunos do curso de Engenharia Elétrica, da Universidade de Eindhover, para predizer a evasão de estudantes. Utilizando classificadores populares da ferramenta Weka, os autores obtiveram taxas de acerto entre 75% e 80%.

Baradwaj e Pal [6] conduziram um estudo para auxiliar professores do curso de Computação, na divisão de alunos em grupos, conforme previsão de desempenho, utilizando um método de classificação bayesiana, com base em dados socioeconômicos e notas do ensino médio de alunos anteriores. Os autores constataram a relação entre esses dados e a nota final obtida pelo aluno no curso. Baradwaj e Pal [5] aplicaram técnicas de DM em dados de alunos de pós-graduação em Computação, com finalidade de predizer as notas do final do semestre. Foram utilizadas informações de frequência, notas de testes, seminários e atividades, de semestres anteriores cursados pelos alunos, como atributos de entrada. Os autores definiram um conjunto de regras, a partir de uma árvore de decisão, para classificar os estudantes de acordo com o desempenho esperado.

Manhães et al. [15] utilizaram as primeiras notas dos alunos calouros do curso de Engenharia Civil, da UFRJ, para previsão daqueles com risco de evasão. Os autores obtiveram acurácias de até 80% na classificação dos alunos. Brito et al. [7] propuseram um método para predição de desempenho dos alunos do primeiro período, do curso de Ciência da Computação, da UFPB, baseado nas notas de ingresso na instituição. Com o uso de classificadores, os autores alcançaram taxas de acerto de até 75%.

3. SELEÇÃO DOS DADOS

Neste estudo, foram utilizados dados do histórico de alunos do curso de Bacharelado em Ciência da Computação, da UFPB. Os dados foram fornecidos pela Superintendência de Tecnologia da Informação (STI) da universidade e contêm registros das notas do vestibular e do histórico de alunos do ano de 2001 até 2013. Os registros apresentam apenas dados de alunos com matriculas inativas, isto é, que concluíram ou abandonaram o curso. O aluno é considerado concluinte se foi aprovado em todas as disciplinas do curso, no prazo máximo determinado (seis anos). Considera-se um aluno evadido caso não efetue matrícula em disciplinas, nem realize trancamento total do curso, durante dos períodos letivos consecutivos, ou caso não tenha concluído o curso no prazo máximo estabelecido.

O curso dispôs de três matrizes curriculares diferentes nos anos compreendidos entre 2001 e 2013, porém, com pequenas

mudanças em relação às disciplinas iniciais do curso, a equiparação dos fluxogramas foi possível. Foram consideradas as cinco disciplinas cursadas no primeiro período, em comum aos três currículos: "Física Aplicada a Computação I", "Cálculo Integral de Diferencial I", "Cálculo Vetorial e Geometria Analítica I", "Introdução à Programação" e "Introdução ao Computador" e as oito matérias presentes na prova de ingresso da instituição até o ano de 2013, conhecida como Vestibular: "Comunicação e Expressão", "Língua estrangeira", "História", "Geografia", "Matemática", "Biologia", "Física" e "Química". Registros de estudantes com ausência de alguma destas informações foram eliminados. Ao todo, restaram registros de 406estudantes. A Tabela 1 mostra a divisão desses estudantes, de acordo com o a sua situação final no curso (classe).

Tabela 1. Divisão de registros por classe.

Classe	Número de instâncias
Evadido	184
Concluinte	222

As notas obtidas pelos estudantes nas disciplinas do vestibular podem variar de 0 a 1000, enquanto as notas das disciplinas do curso podem variar de 0 a 10. O sistema da UFPB considera as notas de alunos reprovados por falta como 0, desta maneira, estes registros foram mantidos. As notas das cinco disciplinas do primeiro semestre do curso foram utilizadas como variáveis de entrada para os algoritmos de classificação e a variável de saída (classe) é a situação final do aluno no curso e só pode assumir dois valores: 'concluinte' ou 'evadido'. Para verificar a eficiência e viabilidade da predição dos estudantes com risco de evasão, com base no desempenho no vestibular e nas disciplinas iniciais, foram realizados testes com algoritmos de classificação, conforme discutido na seção seguinte.

4. TESTES E DISCUSSÃO DE RESULTADOS

A fim de atestar a viabilidade das previsões de evasão dos estudantes, realizaram-se testes de classificação no conjunto apresentado na seção anterior. Fez-se uso de algoritmos populares de aprendizagem de máquina, na ferramenta Weka, que apresenta uma coleção destes algoritmos, além de ferramentas de pré-processamento, permitindo que usuários comparem diferentes métodos e identifiquem os mais adequados ao problema [21]. Para tal, utilizou-se a interface *Explorer*, que permite a utilização de todos os recursos do software facilmente. Cada registro dos estudantes é representado pelo conjunto de atributos (notas individuais) e a sua classe relacionada (concluinte ou evadido).

Primeiramente, verificaram-se quais os atributos mais relevantes para a classificação, quanto ao seu ganho de informação. As notas das disciplinas do primeiro período do curso foram as mais relevantes, na seguinte ordem: "Introdução a Programação", "Cálculo Vetorial e Geometria Analítica I", "Cálculo Integral de Diferencial I", "Física Aplicada a Computação I" e "Introdução ao Computador", seguidas pelas notas das matérias exatas do vestibular: "Física", "Química" e "Matemática" e, por fim, as matérias "Biologia" e "Geografia". As matérias "História", "Língua Estrangeira" e "Comunicação e expressão" não foram relevantes para a classificação dos estudantes quanto a sua situação no curso. Em seguida foi realizada uma etapa de pré-processamento, com a finalidade de melhorar a eficiência dos algoritmos na predição da situação final dos estudantes: Os atributos de entrada (notas) foram discretizados, isto é, divididos em conjuntos finitos de estados. Frequentemente há uma melhora nos resultados dos algoritmos, quando se utilizam atributos discretos, mesmo naqueles capazes de utilizar dados numéricos [21]. A discretização foi feita por meio da aplicação do filtro supervisionado *Discretize*.

Terminada a etapa de pré-processamento, iniciaram-se os testes com os algoritmos. A ferramenta Weka divide os algoritmos de aprendizagem de máquina em oito categorias distintas, sendo uma delas específica para classificação multi-instância, o que não é o nosso propósito. Dessa forma optamos por selecionar o melhor algoritmo de cada uma das sete categorias restantes, alterando os parâmetros dos algoritmos por tentativa e teste, visando encontrar o conjunto de parâmetros que produzisse os melhores resultados. Após a etapa de testes, selecionamos os algoritmos: **NaiveBayes**, do grupo dos classificadores Bayesianos; **J48,** integrante das árvores de classificação; **AdaBoostM1,** pertencente aos algoritmos de Meta-aprendizagem; **SimpleLogistic,** parte das funções matemáticas; **IBk**, membro dos Classificadores Preguiçosos; **DecisionTable**, dos classificadores baseados em regra e **VFI**, da classe de classificadores *Miscellaneous*. Em todos os testes, a divisão entre o conjunto de treinamento e o conjunto de testes foi feita através de Validação Cruzada (*CrossValidation*) com o número de 10 *folds*, padrão na ferramenta Weka. Nesta abordagem os algoritmos são executados 10 vezes, utilizando a proporção de 90% do conjunto para treinamento e 10% para testes. Em cada execução são utilizados segmentos distintos para treinamento e teste [7].

Como medida de desempenho dos algoritmos na tarefa de classificação, utiliza-se o total de instâncias classificadas corretamente (acurácia) e as taxas de estudantes evadidos classificados corretamente e incorretamente (verdadeiros positivos e falsos negativos, respectivamente), e estudantes concluintes classificados corretamente e incorretamente (verdadeiros negativos e falsos positivos, respectivamente). A Tabela 2 exibe as quatro possíveis saídas para o problema de classificação de duas classes, de acordo com o valor real e o predito.

Tabela 2. Saídas possíveis para a classificação.

	Evadido (Predito)	**Concluinte (Predito)**
Evadido (Real)	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Concluinte (Real)	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Os resultados dos testes são mostrados na Tabela 3. Como forma de ampliar a análise dos resultados, utiliza-se a curva de características de operação do receptor (do Inglês, *receiver operating characteristics,* ROC). Esta curva é usada para visualizar, organizar e selecionar classificadores baseados no seu desempenho e é capaz de fornecer uma medida mais rica do desempenho de classificação, se comparado a medidas escalares [9]. O gráfico ROC descreve a relação entre o benefício (verdadeiros positivos) e o custo (falsos positivos). A curva dos algoritmos de melhor desempenho se sobrepõe as dos outros menos eficientes.

Tabela 3. Resultado das Execuções dos Algoritmos.

	NaiveBayes	J48	Ada BoostM1	Simple Logistic	IBk	Decision Table	VFI
Acurácia	85,7%	84,0 %	85,7%	85%	85,9%	86,9%	86,2%
Taxa de VP	82,9%	82,4%	85,1%	83,3%	84,2%	86,5%	82,9%
Taxa de FN	17,1%	17,6%	14,9%	16,7%	15,8%	13,5%	17,1%
Taxa de VN	89,1%	85,9%	86,4%	87,0%	88,0%	87,5%	90,2%
Taxa de FP	10,9%	14,1%	13,6%	13,0%	12,0%	12,5%	9,8%

Observa-se, na Tabela 3, que a acurácia da classificação dos alunos se manteve superior a 84% em todos os algoritmos, sendo o *DecisionTable* o que obteve o melhor resultado (86,9%). A classificação correta dos estudantes com potencial de evasão é mais importante do que a dos prováveis concluintes, já que são estes estudantes que necessitam de suporte docente adequado para que não abandonem a universidade por insucesso nas disciplinas. Desta forma, é fundamental observar as taxas de estudantes evadidos classificados corretamente (verdadeiros positivos) como fator de eficiência dos algoritmos. Para este critério, a menor taxa de acerto na classificação destes estudantes foi de 82,4%, no algoritmo *J48*. Estudantes evadidos foram identificados com taxas de acerto de até 86,5%, no algoritmo *DecisionTable*. A taxa de FP é menos relevante do que a taxa de FN. Mesmo assim, a taxa de FP foi menor em todos os algoritmos, ficando abaixo dos 14,1%, atestando a eficiência do método na detecção dos estudantes concluintes. A Figura 1 mostra as curvas ROC dos métodos de aprendizagem de máquina utilizados nos testes. Também se exibe a curva que representa a estratégia de adivinhação randômica de uma classe, como forma de comparação dos resultados. Na curva ROC, o eixo horizontal representa a especificidade do algoritmo, isto é, a proporção de FP e o eixo vertical representa a sua sensibilidade, que consiste na taxa de VP. Analisando as curvas dos algoritmos, verifica-se que os algoritmos utilizados produziram bons resultados na classificação. Entretanto, a semelhança entre as curvas dificulta a comparação dos diferentes algoritmos, desta forma é interessante dispor de outros meios de comparação.

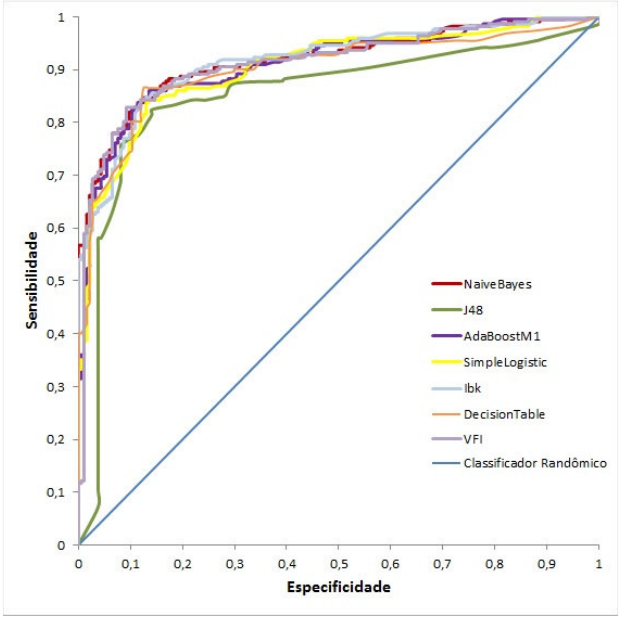

Figura 1. Curvas ROC dos classificadores testados.

Um método comum para comparar diferentes classificadores é utilizar a área sob as curvas ROC (do Inglês, *área under the curve*, AUC). A área varia de 0 a 1, porém deve-se considerar que 0,5 corresponde a AUC do classificador randômico, então valores menores que este não são considerados. Classificadores que possuem maiores AUC apresentam, em média, melhor desempenho na classificação [9]. Em relação às AUC dos classificadores, temos os seguintes resultados (do melhor para o pior): *NaiveBayes* – 0,918; *IBk* – 0,917; *VFI* – 0,913; *AdaBoostM1* – 0,911; *SimpleLogistic* – 0,905; *DecisionTable* – 0,901 e *J48* – 0,853. Para esta métrica, o classificador *NaiveBayes* obteve os melhores resultados. Os classificadores *Ibk* e *AdaBoostM1* tiveram resultados muito próximos e igualmente bons. AUC com valores entre 0,8 e 0,9 indicam excelentes classificações e acima de 0,9 apontam classificações extraordinárias [13]. Desta forma, analisando as AUC dos classificadores, conclui-se que os resultados obtidos pelos

classificadores foram bastante satisfatórios na identificação de estudantes com risco de evasão.

Vale ressaltar que muitas vezes estudantes com bom rendimento optam pela desistência do curso por outros fatores, como falta identificação com a área e falta de perspectiva em relação à carreira profissional no setor [4, 20]. Do mesmo modo, é possível que estudantes com baixo desempenho concluam o curso. A identificação destes estudantes baseada nas suas notas foge ao propósito deste trabalho.

5. CONCLUSÕES E TRABALHOS FUTUROS

A evasão de estudantes é um sério problema enfrentado pelas instituições de ensino superior. A identificação precoce de estudantes com este perfil permite que docentes planejem ações com propósito de evitar o abandono do curso, contribuindo com a diminuição dos altos índices do problema. Neste estudo, propôs-se a aplicação de algoritmos de aprendizagem de máquina na identificação de estudantes com risco de evasão, com base em suas notas nas matérias do vestibular e disciplinas do primeiro semestre do curso. Dados do histórico de alunos do curso de Ciência da Computação, da UFPB, de um período de 12 anos foram utilizados na elaboração do modelo. Os resultados obtidos, com taxas de acerto de até 86,9%, indicam que esta é uma abordagem viável para a detecção de grupos de risco. Desse modo, espera-se fornecer uma ferramenta que auxilie os docentes na identificação de estudantes que necessitam de apoio, contribuindo para a elaboração de políticas de combate à evasão.

Como sugestão de trabalho futuro, acredita-se que é possível estender este estudo para os semestres subsequentes do curso, verificando a eficiência na predição da evasão de estudantes com base no desempenho nas disciplinas destes períodos. Considera-se também a possibilidade de aplicar estudo similar em outros cursos da universidade.

6. REFERÊNCIAS

[1] Baggi, C.A. S. & Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação*. *2*(16), 355-374.

[2] Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*, 112-118.

[3] Baker, R. S. J., Isotani, S. & de Carvalho, A. M. J. B. (2011). Mineração de dados educacionais: Oportunidades para o Brasil.

[4] Barroso, M. F. & Falcão, E. B. (2004). Evasão universitária: O caso do Instituto de Física da UFRJ. *ENCONTRO NACIONAL DE PESQUISA EM ENSINO DE FÍSICA*, *9*, 1-14.

[5] Bhardwaj, B. K. & Pal, S. (2012). Data Mining: prediction for performance improvement using classification.

[6] Bhardwaj, B. K. & Pal, S. (2012). Mining educational data to analyze students' performance.

[7] Brito, D. M., de Almeida Júnior, I. A., Queiroga, E. V. & do Rêgo, T. G. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. In *Anais do Simpósio Brasileiro de Informática na Educação* (Vol. 25, No. 1, pp. 882-890)

[8] Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*

[9] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874

[10] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37A

[11] García, E., Romero, C., Ventura, S., & de Castro, C. (2006). Using rules discovery for the continuous improvement of e-learning courses. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006* (pp. 887-895). Springer Berlin Heidelberg.

[12] Gottardo, E., Kaestner, C., & 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* (Vol. 23, No. 1)

[13] Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

[14] Kotsiantis, S. B., Pierrakeas, C. J. & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 267-274). Springer Berlin Heidelberg.

[15] Manhães, L. M. B., Cruz, S. D., Costa, R. J. M., Zavaleta, J. & Zimbrão, G. (2011). Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. *Anais do XXII SBIE-XVII WIE, Aracaju*.

[16] Romero, C. & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *40*(6), 601-618.

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

[18] Santos, H., Camargo, F. & Camargo, S. (2012). Minerando Dados de Ambientes Virtuais de Aprendizagem para Predição de Desempenho de Estudantes. *Conferencias LACLO*, *3*(1).

[19] Seidman, A. (Ed.). (2005). *College student retention: Formula for student success*. Greenwood Publishing Group.

[20] Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., & Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*,*37*(132), 641-659.

[21] Witten, I. H., & Frank, E. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. 3ª Edição.