

Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais.

Ernani Gottardo¹, Celso Kaestner², Robinson Vida Noronha²

¹Instituto Federal de Educação Ciência e Tecnologia do RS (IFRS)
CEP 99.700-000 – Erechim – RS – Brasil

²Universidade Tecnológica Federal do Paraná (UTFPR)
CEP 80.230-901 – Curitiba – PR – Brasil.

ernani.gottardo@erechim.ifrs.edu.br, {celsokaestner,vida}@utfpr.edu.br

Abstract. *Providing relevant information to assist development of effective processes for monitoring students in distance learning courses is one of the goals of scientific community on Computers in Education. In this work, Educational Data Mining techniques were used to generate inferences about the performance of students based on data gathered in time series. The main objective is to investigate the viability of obtaining this information in the initial stages of the course in order to support the adoption of proactive actions. The results show that these inferences can be obtained with accuracy rates close to 75%, even in the early periods of the course.*

Resumo. *Prover informações relevantes que auxiliem o desenvolvimento de processos de acompanhamento efetivo de estudantes em cursos a distância é um dos objetivos da comunidade científica de Informática na Educação. Neste trabalho, técnicas de mineração de dados educacionais foram utilizadas para geração de inferências sobre o desempenho de estudantes a partir de dados coletados em séries temporais. O objetivo principal é investigar a viabilidade de obtenção destas informações em etapas iniciais de realização do curso de forma a apoiar a tomada de ações pró-ativas. Os resultados obtidos demonstram que é possível obter estas inferências com taxas de precisão próximas a 75%, mesmo em períodos iniciais do curso.*

1. Introdução

O desenvolvimento e operacionalização de processos de acompanhamento efetivo dos estudantes em cursos a distância tem se mostrado um desafio para professores, instituições e profissionais que atuam com esta modalidade de ensino.

Esta situação torna-se relevante tendo em vista que um fator importante para a efetividade de experiências de aprendizagem está relacionada com a capacidade de professores em monitorar este processo e tomar decisões com base em eventos observados [Moore 2007; Romero-Zaldivar et al. 2012]. Por exemplo, a informação de que um recurso disponibilizado em um Ambiente Virtual de Aprendizagem - AVA não foi acessado pelos estudantes conforme esperado poderia indicar que: I) o professor não deixou claro que o conteúdo deveria ser estudado (instruções inadequadas) ou II) o

conteúdo não está sendo encontrado pelos estudantes (problema de visibilidade) [Ricarte e Falci Junior 2011].

O uso de ferramentas tecnológicas que forneçam suporte à obtenção e disponibilização de informações relativas ao processo de aprendizagem poderiam auxiliar a enfrentar esses desafios [Bittencourt e Costa 2011]. Essas ferramentas têm o potencial de fornecer informações úteis para subsidiar o desenvolvimento de ações (e.g. diminuir a reprovação ou evasão, melhoria de qualidade do curso) ou estratégias pedagógicas específicas. Essas informações poderiam também evidenciar algumas características individuais ou coletivas de estudantes.

Neste sentido, a comunidade de pesquisa em Informática na Educação tem investigado as possibilidades de utilização de banco de dados com grandes volumes de informações sobre estudantes gerados pelo uso de softwares de apoio à atividade educacional, como os Ambientes Virtuais de Aprendizagem (AVAs).

As informações armazenadas nestes bancos de dados são bastante amplas. Elas incluem desde simples registros de acesso e interações com o sistema, até dados com ricos significados semânticos, como respostas a testes ou participações em fóruns e chats [Romero e Ventura 2010; Romero-Zaldivar et al. 2012]. Entretanto, identificar quais dessas informações são relevantes e manipulá-las de maneira significativa não é um tarefa simples.

Este trabalho apresenta um estudo preliminar do uso de algumas técnicas de Mineração de Dados Educacionais em uma base de dados do AVA Moodle. O objetivo principal é a obtenção de um modelo de previsão de desempenho de estudantes a partir de dados coletados em séries temporais durante o período de realização de um curso a distância. Com esse modelo, pretende-se realizar inferências sobre o desempenho dos estudantes e disponibilizá-las a professores em etapas iniciais de realização do curso.

Na descrição dos resultados deste trabalho, este documento está estruturado em 6 seções. Na Seção 2, analisam-se os trabalhos correlatos. Na seção 3, apresentam-se os critérios utilizados para seleção e tratamento dos dados. Na seção 4 e 5, são apresentadas, respectivamente, a descrição dos experimentos realizados e a análise dos resultados obtidos. Finalmente, na seção 6, destacam-se as conclusões e perspectivas de continuidade deste trabalho.

2. Trabalhos Relacionados

Uma linha de pesquisa que apresenta potencial para melhorar a qualidade do ensino e que tem se consolidado nos últimos anos é conhecida como Mineração de Dados Educacionais ou EDM [Baker, Isotani e Carvalho 2011]. EDM é a área de investigação científica centrada no desenvolvimento de métodos para realizar descobertas de conhecimento em base de dados educacionais. Para atingir os objetivos propostos, os métodos utilizados nesta linha de pesquisa exploram, em conjunto ou individualmente, a aplicação de técnicas como Mineração de Dados (*Data Mining*), estatísticas e aprendizado de máquina [Romero e Ventura 2010].

Mineração de Dados é definida por Witten, Frank e Hall (2011) como o processo de descoberta de padrões a partir de um conjunto de dados. Fayyad, Piatetsky-Shapiro e Smyth (1996), completam que os padrões descobertos precisam revelar alguma

novidade e serem potencialmente úteis de forma a trazer algum benefício para o usuário ou auxiliar o desenvolvimento de alguma atividade.

No contexto educacional, espera-se que as possíveis descobertas realizadas tenham potencial de auxiliar no entendimento de como os alunos aprendem e os principais fatores que causam impactos no processo de aprendizagem.

Baker, Isotani e Carvalho (2011) citam algumas possibilidades de descobertas importantes, como por exemplo: i) identificar em que situação um tipo de abordagem instrucional (e.g. aprendizagem individual ou colaborativa) proporciona melhores benefícios educacionais ao aluno e ii) verificar se um aluno está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem.

Tomando-se como base as informações disponíveis sobre cursos, dados de uso e interação de estudantes, diferentes técnicas de Mineração de Dados podem ser aplicadas. Diante disso, Romero, Ventura e García (2008) apresentaram um estudo apontando as possibilidades de utilização de algumas destas técnicas, tais como: classificação, agrupamento, regras de associação e mineração de texto (*Text Mining*).

A técnica de Mineração de Dados utilizada neste trabalho é conhecida como classificação. Através desta técnica busca-se obter modelos para inferir um aspecto particular dos dados (variável a ser prevista) através de alguma combinação de outros aspectos destes dados (variáveis preditoras) [Witten, Frank e Hall 2011]. No presente trabalho a variável a ser prevista é o desempenho dos estudantes, utilizando-se para tal um conjunto de variáveis preditoras extraídas de um banco de dados de um AVA.

Baseando-se na concepção de que o algoritmo de classificação ideal é altamente dependente do domínio do problema e com o objetivo de melhorar o desempenho de algoritmos aplicados a área educacional, Minaei-Bidgoli et al. (2003) demonstraram a aplicação de uma técnica promissora chamada de CMC (*Combination of Multiple Classifiers*). Os autores concluíram que esta técnica apresenta melhores resultados quando comparada a aplicação de classificadores individualmente.

Alguns trabalhos envolvendo aplicação de técnicas de EDM foram desenvolvidos com o objetivo de comparar o desempenho de diversas técnicas e algoritmos de Mineração de Dados aplicadas a problemas de previsão de desempenho [Ibrahim e Rusli 2007; Kotsiantis 2012; Romero-Zaldivar et al. 2012] e risco de evasão [Manhães et al. 2011]. Os resultados das referidas pesquisas demonstram que diversas técnicas de mineração têm potencial de serem utilizadas no contexto educacional para previsão de desempenho ou risco de evasão.

Com o objetivo de fornecer *feedback* para autores de materiais e tutores a respeito do uso dos materiais disponíveis em um curso, Ricarte e Falci Junior (2011) desenvolveram um estudo utilizando algoritmos de agrupamento. Com a aplicação desta técnica em cursos utilizando um AVA na UNICAMP, os autores buscaram descobrir grupos de estudantes com comportamento similar e então definir padrões de uso e identificar estudantes com comportamento inadequado.

Considerando que os ambientes clássicos de educação a distância (AVA) apresentam limitações relativas à falta de controle e avaliação personalizada, além de deficiências na adaptação às características dos usuários, Bittencourt e Costa (2011) investigam soluções e metodologias para o desenvolvimento de Ambientes

Educacionais Adaptativos e Semânticos. A principal característica de ambientes educacionais adaptativos e semânticos refere-se à preocupação com a automatização, interoperabilidade e reuso entre aplicações. Com isso busca-se prover adaptação aos usuários e integração com ferramentas para descrição semântica do domínio, de modo a garantir um processamento automático [Bittencourt e Costa 2011].

Visando o desenvolvimento de melhores Sistemas Tutores Inteligentes (ITS), métodos de EDM estão sendo aplicados na busca de evidências para refinar e estender teorias e fenômenos educacionais conhecidos.

Por sua vez, outras pesquisas têm demonstrado o uso de métodos de EDM para inferir a relação existente entre o estado emocional de estudantes e o desempenho apresentado por eles [Mcquiggan, Mott e Lester 2008] ou para identificar quando um estudante está tentando manipular o sistema, em inglês “*gaming the system*” [Baker et al. 2006].

Levando-se em consideração a natureza social do aprendizado, algumas pesquisas [Macfadyen e Dawson 2010; Rabbany, Takaffoli e Zaiane 2011] têm focado no estudo de como estudantes participam e interagem em um ambiente de aprendizagem virtual. Dawson e Mcwillian (2008) demonstram como a técnica conhecida como Análise de Redes Sociais pode ser aplicada para extrair registros de interação a partir de fóruns de discussão e analisar algumas medidas que representam aspectos das relações entre os estudantes.

Em resumo, diversos trabalhos investigaram a aplicação de técnicas de Mineração de Dados Educacionais com o objetivo de realizar inferências ou previsão a respeito do desempenho de estudantes. Entretanto, de forma geral, são analisados os resultados considerando dados envolvendo todo o período de realização do curso, que só podem ser obtidos ao final do mesmo. Desta forma, as informações obtidas, embora consideradas importantes ao contexto educacional, somente poderão ser utilizadas como apoio para ações envolvendo estudantes de cursos futuros.

3. Seleção e Tratamento dos Dados

Para a realização deste trabalho, utilizou-se uma base de dados do AVA Moodle com informações sobre cursos realizados a distância. Do ambiente Moodle, selecionou-se uma disciplina que atendesse os seguintes requisitos experimentais: i) maior quantidade de estudantes que concluíram a disciplina, ii) maior quantidade de oferta da mesma disciplina para turmas diferentes, iii) disponibilidade do resultado de avaliações do conhecimento dos estudantes e iv) maior número de recursos do AVA utilizados.

Com base nos critérios apresentados acima, escolheu-se uma disciplina com um total de 140 estudantes concluintes em duas turmas diferentes. Não foram considerados estudantes desistentes dessas turmas, tendo em vista a indisponibilidade de resultado ou nota final nesses casos.

A escolha dos atributos para compor o conjunto de dados experimental sobre os 140 estudantes selecionados foi baseada no modelo de representação de estudantes proposto por Gottardo, Kaestner e Noronha (2012). A Tabela 1 reapresenta esses atributos, agrupados nas três dimensões propostas no referido trabalho, juntamente com uma breve descrição de cada atributo. Considerando os atributos apresentados nessa tabela, foi desenvolvido procedimento para extração de dados do AVA Moodle.

Tabela 1. Atributos propostos para representação de estudantes

Dimensão	Atributo	Descrição
Perfil Geral de uso do AVA	nr_acessos	Número total de acesso ao AVA
	nr_posts_foruns	Número total de postagens em fóruns
	nr_post_resp_foruns	Número total de respostas postadas em fóruns referindo-se a postagens de outros participantes
	nr_post_rev_foruns	Número total de revisões em postagens anteriores realizadas em fóruns
	nr_sessao_chat	Número de sessões de chat que o estudante participou
	nr_msg_env_chat	Número de mensagens enviadas ao chat
	nr_questoes_resp	Número de questões respondidas
	nr_questoes_acert	Número de questões respondidas corretamente
	freq_media_acesso	Frequência que o estudante acessa o AVA
	tempo_medio_acesso	Tempo médio de acesso ao sistema
	nr_dias_prim_acesso	Número de dias transcorridos entre o início do curso e o primeiro acesso ao AVA
	tempo_total_acesso	Tempo total conectado no sistema
Interação Estudante-Estudante	nr_post_rec_foruns	Número de postagens do estudante que tiveram respostas feitas por outros estudantes.
	nr_post_resp_foruns	Número de respostas que o estudante realizou em postagens feitas por outros estudantes.
	nr_msg_rec	Número de mensagens recebidas de outros participantes durante a realização do curso.
	nr_msg_env	Número de mensagens enviadas a outros participantes durante a realização do curso.
Interação Estudante-Professor	nr_post_resp_prof_foruns	Número de postagens de estudantes que tiveram respostas feitas por professores
	nr_post_env_prof_foruns	Número de postagens de professores que tiveram respostas feitas por estudantes
	nr_msg_env_prof	Número de mensagens enviadas ao professor/tutor durante a realização do curso.
	nr_msg_rec_prof	Número de mensagens recebidas do professor/tutor durante a realização do curso.
Objetivo da Previsão	resultado_final	Nota final obtida pelo estudante. <u>Representa classe objetivo da técnica de classificação.</u>

Fonte: Gottardo, Kaestner e Noronha (2012)

O atributo “resultado_final” apresentado na Tabela 1 armazena a nota final obtida pelo estudante e representa a variável a ser prevista pelo processo de classificação. Esse atributo é definido originalmente para armazenar valores contínuos. Essa particularidade representa uma limitação em alguns procedimentos classificatórios. Algumas técnicas de classificação requerem que o atributo que representa a classe seja um valor discreto e não contínuo. Além disso, alguns algoritmos de Mineração de

Dados, mesmo os capazes de trabalhar com valores contínuos, podem apresentar melhores resultados quando se utilizam valores discretos [Witten, Frank e Hall 2011].

Diante disso, para viabilizar a utilização de alguns tipos de métodos e também para facilitar a interpretação dos resultados os valores do atributo “resultado_final” foram transformados em valores discretos, conforme procedimento descrito a seguir.

Analizando-se as notas dos estudantes das turmas selecionadas pode-se observar que essas seguem uma distribuição normal. Nestas turmas, a média das notas dos estudantes foi 82, com desvio padrão igual a 5. A menor nota foi 67 e a maior 97.

Optou-se por dividir os estudantes em três classes de acordo com a nota obtida no curso. Inicialmente, definiu-se uma classe intitulada como “B” para agrupar os estudantes com desempenho intermediário. Nesta classe foram alocados os estudantes cuja probabilidade das notas situem-se em aproximadamente 80% da área central da curva normal. Os demais estudantes foram distribuídos nas classes “C” e “A”, representando, respectivamente, aproximadamente 10% da área inferior e superior.

A Figura 1 apresenta graficamente a distribuição das classes obtidas com a aplicação do processo descrito acima. A região central desta figura (destacada em cinza) compreende aproximadamente 80% da área total e representa a classe B. Nas extremidades da curva (cor branca) localizam-se os estudantes das classes C e A. Os números apresentados na horizontal representam as notas que delimitam as classes, com o valor médio 82 ao centro.

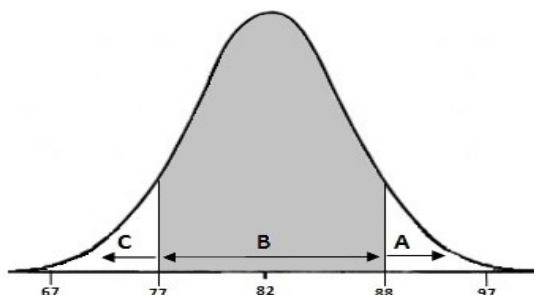


Figura 1- Visualização da distribuição das classes

A Tabela 2 apresenta mais detalhes sobre os resultados do processo de discretização descrito acima, destacando algumas das características individuais de cada classe. Nesta tabela são apresentados o título da classe, a descrição, o intervalo de notas e o número de estudantes distribuídos em cada uma das classes.

Tabela 2. Distribuição das classes obtidas pelo processo de discretização

Título da Classe	Descrição	Intervalo de Notas	Número de Estudantes
A	Estudantes com desempenho superior	Maior que 88	16
B	Estudantes com desempenho intermediário	Entre 77 e 88	109
C	Estudantes com desempenho inferior	Menor que 77	15

É importante ressaltar que existem diversas alternativas para o processo de discretização apresentado acima. Além disso, o número de classes e o número de instâncias em cada classe podem variar em função dos objetivos e das características particulares de cada estudo. Essas definições são relevantes, pois podem influenciar os

resultados obtidos pelos modelos de previsão.

4. Experimentos Realizados

Tomando-se como base o conjunto de dados com os atributos descritos na Tabela 1, dois experimentos foram realizados. Esses experimentos avaliam as possibilidades de obtenção de inferências sobre o desempenho de estudantes em diferentes etapas de realização do curso.

A medida utilizada para avaliação dos resultados dos experimentos é a Acurácia. Define-se Acurácia à proporção entre o número de estudantes corretamente classificados pelos algoritmos em sua respectiva classe e o número total de estudantes considerados no estudo.

Para o cálculo da acurácia utilizou-se o método conhecido como “K - fold Cross-Validation” [Witten, Frank e Hall 2011] que consiste de uma técnica para a estratificação da base dados em conjunto de treinamento e teste. Adicionalmente, cada algoritmo foi executado 10 vezes e considerou-se a média aritmética dessas 10 execuções como o desempenho final do algoritmo. Logo, em cada experimento o classificador foi executado 100 vezes para cada conjunto de treinamento e testes.

Nos experimentos realizados foram utilizados os algoritmos de classificação “RandomForest” e “MultilayerPerceptron” [Witten, Frank e Hall 2011]. Esta escolha baseou-se nas conclusões apresentadas por alguns trabalhos correlatos destacados na Seção 2 que demonstraram que estes dois algoritmos obtiveram os melhores resultados.

No experimento 1, o período total de realização do curso foi dividido em três, considerando-se o número de módulos em que o curso em estudo é dividido. Foram coletados três conjuntos de dados distintos, cada um contendo informações compreendendo a data de início do curso até a data final do módulo em questão.

A Tabela 3 demonstra os resultados desse experimento em cada um dos três módulos, apresentando o percentual de acurácia médio e o desvio padrão de 100 execuções de cada algoritmo. A Tabela 3 mostra que, por exemplo, o algoritmo RandomForest, utilizando o conjunto de dado do Módulo 1, classificou aproximadamente 107 (76.2% \pm 4.7) dos 140 estudantes na classe correta (A,B ou C).

Tabela 3. Acurácia média e desvio padrão em 100 execuções dos classificadores utilizados no experimento 1

Classificador	Módulo 1	Módulo 2	Módulo 3
RandomForest	76.2% \pm 4.7	77.8% \pm 2.1	77.2% \pm 2.9
MultilayerPerceptron	76.5% \pm 4.4	77.8% \pm 2.1	77.2% \pm 2.9

O experimento 2 foi realizado dividindo-se o tempo total de realização do curso em seis períodos iguais. A divisão dos períodos não levou em consideração datas de início e final de módulos ou disciplinas.

A Tabela 4 demonstra os resultados deste experimento em cada um dos seis períodos, apresentando o percentual de acurácia médio e o desvio padrão de 100 execuções para os dois algoritmos utilizados. A Tabela 4 mostra que, por exemplo, o algoritmo MultilayerPerceptron, utilizando o conjunto de dado do Período 1, classificou aproximadamente 104 (74.4% \pm 6.2) dos 140 estudantes na classe correta (A,B ou C).

Tabela 4. Acurácia média e desvio padrão em 100 execuções dos classificadores

Classificador	Períodos de Tempo					
	1	2	3	4	5	6
RandomForest	75.2% ±5.7	75.2% ±7.0	81.5% ±5.9	77.2% ±2.9	77.2% ±2.9	77.2% ±2.9
MultilayerPerceptron	74.4% ±6.2	74.0% ±6.7	81.3% ±6.1	77.2% ±2.9	77.2% ±2.9	77.2% ±2.9

5. Análise dos Resultados

Uma avaliação inicial dos resultados obtidos e descritos nas Tabelas 3 e 4 demonstra que os percentuais de acurácia não têm grande variação entre os diversos pontos de corte utilizados: módulos do experimento 1 e períodos de tempo do experimento 2. Na Tabela 3, a maior diferença observada é de 1,6% entre o Módulo 1 e 2 para o algoritmo RandomForest. Na Tabela 4, a maior diferença entre os percentuais de acurácia é 6,9% observada entre o Período 1 e 3 quando utilizado o algoritmo MultilayerPerceptron.

Para testar a significância estatística dos resultados, utilizou-se a técnica “T pareado” – “*pair-wise T-Test*” [Witten, Frank e Hall 2011] com nível de significância de 5%. Este teste foi realizado utilizando-se o ambiente *Weka Experiment Environment* – WEE. A partir do resultado deste teste, tomando-se como base ou *baseline* o módulo 1 no experimento 1 e o período 1 do experimento 2, observou-se que:

- As taxas de acurácia obtidas no experimento 1 não apresentam diferença estatisticamente significativa entre o módulo 1 e os demais módulos. Este resultado repete-se nos dois algoritmos utilizados.
- No experimento 2, apenas o período 3 apresenta diferença estatisticamente significativa em relação ao período 1. Em ambos algoritmos utilizados no experimento 2, este resultado repete-se.

Nos dois experimentos o maior percentual de acurácia foi obtido na parte intermediária das divisões realizadas. Entretanto, apenas no experimento 2 a taxa obtida neste período foi estatisticamente superior ao período inicial. A explicação deste fato poderia ser investigada através de uma análise mais detalhada da organização, estrutura e atividades desenvolvidas, preferencialmente envolvendo pessoas ligadas ao curso, como coordenadores, professores, tutores, entre outros.

Observa-se também que mesmo a menor taxa de acurácia obtida (74,4% alcançada pelo algoritmo MultilayerPerceptron no Experimento 2, conforme Tabela 4) pode ser considerada relevante no contexto educacional. Esta afirmação baseia-se no resultado de Hämäläinen e Vinni (2011) que, utilizando pesquisas envolvendo previsão de desempenho, realizaram um estudo para verificar as taxas de acurácia obtidas. Os referidos autores destacam que devido às características particulares dos estudos existem variações entre os resultados observados. Porém, destacam que a média de acurácia foi de 72% entre as pesquisas consideradas.

6. Considerações Finais e Trabalhos Futuros

A disponibilização de informações e ferramentas que permitam o efetivo acompanhamento de estudantes em cursos EAD é considerada pela comunidade científica de Informática na Educação como um desafio ainda em aberto. Neste trabalho investigou-se a viabilidade de aplicação de técnicas de mineração de dados para realizar inferências relativas ao desempenho de estudantes em diferentes fases ou etapas de realização de um curso. Essas inferências consistem em classificar previamente (ao final de um módulo ou período) os estudantes em três classes (A, B e C) de acordo com a previsão de desempenho dos mesmos ao final do curso.

Os resultados obtidos com os experimentos apontam para viabilidade de se realizar inferências relativas ao desempenho de estudantes, obtendo-se taxas de acurácia acima de 74%, mesmo em etapas iniciais de realização do curso. Estimativas sobre o desempenho, disponibilizadas durante a realização do curso, poderiam ser úteis para professores acompanharem individualmente estudantes e desenvolverem estratégias pedagógicas que busquem minimizar reprovações. Essas informações disponibilizadas em períodos iniciais do curso poderiam ser utilizadas para o desenvolvimento de ações que envolvam estudantes de turmas em andamento e não somente para futuras turmas.

A metodologia para a definição das classes poderia ser tratada como tema de trabalhos futuros, visando investigar o impacto de mudanças nesse procedimento nos resultados dos modelos de inferência de desempenho. Além disso, futuras pesquisas poderiam avaliar técnicas aplicáveis ao contexto educacional para tratar a questão do desbalanceamento entre o número de instâncias das classes, visto que isso pode interferir significativamente no resultado de classificadores.

Referências

- Baker, R.S.J.D., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006). "Adapting to When Students Game an Intelligent Tutoring System". In Proceedings of the International Conference on Intelligent Tutoring Systems, p. 392-401.
- Baker, R.S.J.D., Isotani, S., Carvalho, A.M.J.B.D. (2011) "Mineração de Dados Educacionais: Oportunidades para o Brasil". Revista Brasileira de Informática na Educação, vol. 19, no. 2, p. 2-13.
- Bittencourt, I. I., Costa, E.B. (2011) "Modelos e Ferramentas para a Construção de Sistemas Educacionais Adaptativos e Semânticos". Revista Brasileira de Informática na Educação, v. 19, p. 85-98.
- Dawson, S., McWilliam, E. (2008) "Investigating the Application of IT Generated Data as an Indicator of Learning and Teaching Performance". University of British Columbia.
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996) "The KDD Process for Extracting Useful Knowledge from Volumes of Data". Communications of ACM, vol. 39, no. 11, p. 27-34.
- Gottardo, E., Kaestner, C., Noronha, R.V. (2012) "Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados". Anais do XXXII Congresso da Sociedade Brasileira de Computação.

- 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.
- Ibrahim, Z., Rusli, D. (2007) "Predicting students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression", In Proceedings of the 21º Annual SAS Malaysia Forum, Kuala Lumpur, Malaysia, p. 1-6.
- Kotsiantis, S.B. (2012) "Use of Machine Learning Techniques for Educational Proposes: A Decision Support System for Forecasting Student's Grades". Artificial Intelligence Review, vol.37, no. 4, p. 331-344.
- 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.
- Manhães, L.M.B., Cruz, S.M.S., 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, p. 150-159.
- Mcquiggan, S., Mott, B., Lester, J. (2008) "Modeling Self-Efficacy in Intelligent Tutoring Systems: An Inductive Approach". User Modeling and User-Adapted Interaction, vol.18, p. 81-123.
- Minaei-Bidgoli, B., Kashy, A.D., Kortemeyer, G., Punch, F.W. (2003) "Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA". 33ª ASEE/IEEE Frontiers in Education Conference, p. 1-6.
- Moore, M. G. "Three Types of Interaction". (1989) The American Journal of Distance Education, Vol. 3, No. 2, p. 1-6.
- Rabbany, R.K., Takaffoli M., Zaiane, O.R. (2011) "Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques". In Proceedings of the Fourth International Conference on Educational Data Mining, p. 22-30.
- Ricarte, I. L. M., Falci Junior, G. R. (2011) "A Methodology for Mining Data from Computer-Supported Learning Environments". Informática na Educação: teoria & prática, Porto Alegre, v. 14, n. 2, p. 83-94.
- Romero, C. Ventura, S., García, E. (2008) "Data Mining in Course Management Systems: Moodle Case Study and Tutorial". Computers & Education, no. 51, p. 368-384.
- Romero, C. Ventura, S. (2010) "Educational Data Mining: A Review of the State of Art". IEEE Transactions on Systems , Man, and Cybernetics – Part C: Application and Reviews. vol. 40, no. 6, p. 601-618.
- Romero-Zaldivar, V.A., Pardo, A., Burgos, D., Kloos, C.D. (2012) "Monitoring Student Progress Using Virtual Appliances: A Case Study". Computers & Education, no. 58, p. 1058-1067.
- Witten, I.H., Frank E., Hall, M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 3 ed.