Mineração de Dados Educacionais no Apoio ao Processo de Tomada de Decisão do Docente

Alana M. Morais[1], Joseana Fechine[1]

[1]Universidade Federal de Campina Grande (UFCG) – Campina Grande – PB – Brasil

alanamorais@copin.ufcg.edu.br, joseana@dsc.ufcg.edu.br

Abstract. *This paper presents an approach to identify which factors are most relevant in e-learning student's performance. So, we use classification algorithms on study of relevant factors in the sample. The results showed that the score of student's participation and good performance in your tasks influence the positive student's results in the course.*

Resumo. *Este artigo discute uma abordagem de análise para identificar quais são os fatores mais relevantes em um curso de Educação a Distância. Para tanto, foram utilizadas algoritmos de classificação utilizados em Mineração de Dados Educacionais no estudo desses fatores. Os resultados obtidos indicaram que de fato a taxa de participação do aluno e o seu bom desempenho nas atividades cadastradas no ambiente influenciam nos bons resultados desse aluno ao final do curso.*

1. Introdução

As tecnologias digitais estão cada vez mais incorporadas à sala de aula, seja no apoio à modalidade presencial ou na Educação a Distância (EaD), além de prover suporte aos espaços pedagógicos de produção do conhecimento. Neste sentido, destacam-se sistemas computacionais muito comentados na literatura: os Ambientes Virtuais de Aprendizagem (AVA). Estes sistemas precisam, cada vez mais, lidar com um volume maior de dados, possibilitar uma interação mais intuitiva entre o aluno e o ambiente, oferecer novos recursos didático-pedagógicos e prover ferramentas de qualidade para o acompanhamento e avaliação do discente.

Nesse contexto, o estudo ora apresentado tem o intuito de auxiliar o processo de tomada de decisão de professores diante de cursos de EaD, propondo uma abordagem de análise dos dados baseada na aplicação de técnicas para classificação de dados educacionais. As técnicas de classificação foram selecionadas para auxiliar o professor na identificação das relações e da hierarquia existente entre os critérios do ambiente educacional. Para tanto, a metodologia adotada foi dividida em quatro etapas fundamentais: (i) coleta de dados, (ii) análise exploratória, (iii) classificação dos dados e (iv) interpretação dos resultados. Por fim, para validar a abordagem proposta foram utilizados dados obtidos em um estudo de caso. Essa metodologia analisa quais fatores influenciam mais fortemente no desempenho do aluno ao longo da sua interação com o ambiente.

A próxima seção do trabalho faz uma revisão da literatura no âmbito das pesquisas relacionadas ao processo de tomada de decisão do professor nos AVA. Na terceira seção, é discutida a abordagem de análise proposta e na quarta seção a validação dessa abordagem no estudo de caso proposto. Por fim, na quinta seção são discutidas as considerações finais e os trabalhos futuros provenientes dessa análise.

2. Revisão de Literatura

Há poucos sistemas adaptativos para monitorar e controlar as atividades realizadas pelo aluno de EaD, que sejam capazes de detectar e resolver problemas. Algumas pesquisas propõem minimizar essa rigidez aplicando técnicas de Inteligência Artificial (IA) nos ambientes e, com isso, tornando-os adaptativos. Diante de tais técnicas, destaca-se o estudo de Corrigan (2012), que analisa a implantação de sistemas de tutoria virtual em ambientes de EaD frente a um contexto de melhoria na taxa de aprovação dos alunos e limitação do orçamento escolar (AMORIM *et al.*, 2011).

Merece destaque também o processo de acompanhamento do desempenho e participação dos alunos matriculados em um curso, que ocorre por meio de relatórios estáticos gerados pelo próprio ambiente para cada atividade. Estes sistemas oferecem ferramentas para geração de relatório que, em geral, mostram dados brutos (número de acessos, o tempo gasto no curso, etc.) em um formato de tabela. Como consequência, o processo de acompanhamento da progressão acadêmica do discente durante o curso pode se tornar difícil e demandar tempo significativo. Neste sentido, uma técnica que emerge na solução dessas demandas é a Mineração de Dados, a ser discutida na próxima seção.

2.1. Mineração de Dados Educacionais (MDE)

No processo de coleta, análise e interpretação de dados educacionais se destacam as técnicas associadas à Mineração de Dados Educacionais (MDE). Entende-se por MDE o processo de conversão de dados brutos de sistemas educacionais para informações úteis que podem ser usadas para fornecer decisões de projeto e responder a questões de pesquisa (BAKER, 2011).

O trabalho em tela baseou-se na classificação das técnicas de MDE sugeridas por Baker (2011), que destaca os seguintes grupos: predição, agrupamento, relações de mineração, descoberta com modelos e extração de dados pelo julgamento humano. No âmbito do trabalho, são utilizadas as técnicas de predição para extrair informações úteis ao professor, especificamente as técnicas de classificação.

3. Metodologia do Experimento

Nesta seção do artigo, é discutido de forma detalhada o *design* do estudo de caso realizado e a metodologia da abordagem de MDE proposta. Este estudo tem como objetivo obter informações que auxiliem os processos decisórios do professor em um AVA, diante de situações de desempenho inadequado ou risco de desistência do aluno. O trabalho possui duas questões centrais que se propõe a responder, a saber: Quais elementos do AVA podem influenciar no desempenho do aluno? Foi possível extrair algum comportamento, que não havia sido detectado anteriormente nas interações convencionais com o ambiente?

3.1. Abordagem proposta

A metodologia adotada na abordagem proposta é baseada em quatro etapas principais: (i) coleta de dados, (ii) análise exploratória, (iii) classificação dos dados e (iv) interpretação dos resultados, sequência ilustrada na Figura 1. É importante comentar que durante o tratamento dos dados coletados, na primeira etapa da abordagem, ocorrem as primeiras ações de pré-processamento desses. O intuito do pré-processamento, dentre outros, é suprir valores ausentes, reduzir discrepâncias de valores ruidosos e corrigir inconsistências na amostra coletada.

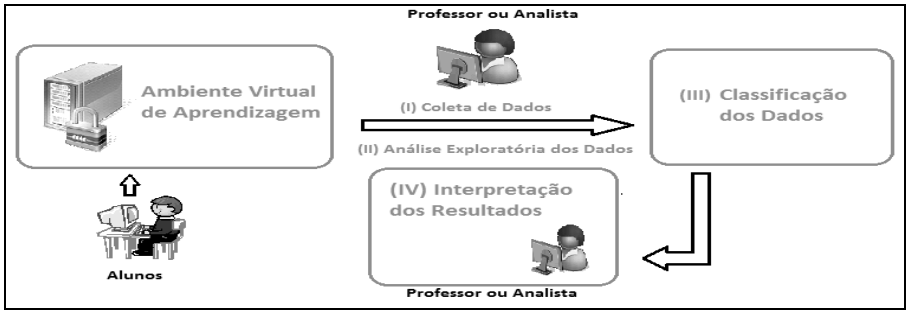

Figura 1. Etapas da abordagem proposta.

3.1.1. *Design* do Experimento

Para validar a abordagem definida e elucidar melhor as etapas comentadas, foram utilizados dados obtidos por meio de um estudo de caso. Os dados foram coletados a partir de cursos no *Moodle* e por isso as informações foram obtidas por meio de relatórios gerados pelo ambiente.

Características da Amostra: Os dados do estudo de caso foram coletados nas disciplinas de Comércio Eletrônico (CE) e de Interface Homem Computador (IHC) nos semestres 2011.2 e 2012.1 no UNIPÊ, sendo composta por 79 alunos. Para as quatro turmas observadas foram utilizadas a mesma metodologia pedagógica de planejamento dos conteúdos e atividades. A cada semana um novo tópico era ativado no ambiente; contendo uma atividade de fórum, bate-papo e exercício a ser entregue via ambiente.

3.2. Técnicas para Classificação dos Dados

A abordagem proposta tem o intuito de prover informações relevantes aos professores aplicando técnicas de IA e Estatística no processo. Para tanto, são utilizadas diversas metodologias e algoritmos de classificação. Neste trabalho, a decisão caracteriza a conclusão sobre o *Status* do discente ao final do curso. O *Status* pode assumir os seguintes valores: "Reprovado Desistente", "Aprovado" e "Reprovado Nota".

a) Árvores de Decisão

Uma das técnicas empregadas na análise de dados e em processos de tomada de decisão rápida são as árvores de decisão (CHIKALOV, 2011). Estas podem ser definidas como modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas. Uma árvore de decisão recebe como entrada um objeto ou situação descrito por um conjunto de atributos e retorna uma decisão - o valor da saída previsto, de acordo com a entrada (RUSSELL e NORVIG, 2004). Ao final do processo de classificação, a árvore de decisão apresentará no seu topo a variável mais influente para as saídas analisadas. Há diversos métodos para geração de árvores de decisão e, no contexto do trabalho, foram utilizados os seguintes métodos: *Id3*, *J48*, *BFTree* e *SimpleCart*.

b) Redes Bayesianas

As redes bayesianas são representações bem desenvolvidas para o conhecimento incerto e desempenham um papel aproximadamente análogo ao da lógica proposicional para o conhecimento definido. Uma rede bayesiana é um grafo acíclico orientado cujos nós correspondem a variáveis aleatórias; cada nó tem uma distribuição condicional para o nó, dados seus pais (RUSSELL e NORVIG, 2004). Outra característica importante é

que as redes bayesianas fornecem um modo conciso de representar relacionamentos de independência condicional no domínio.

Os classificadores bayesianos são ferramentas estatísticas que classificam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe. O objetivo da classificação é prever corretamente o valor de uma variável de classe designado discreto dado um vetor de atributos. A aprendizagem de uma rede bayesiana envolve dois passos principais: primeiro aprende a estrutura da rede e depois as tabelas de probabilidades.

4. Apresentação e Análise dos Resultados Obtidos

Nesta seção são apresentados e discutidos os resultados da aplicação da abordagem no estudo de caso proposto.

4.1. Coleta e Análise Exploratória dos dados

Os primeiros esforços para compreender o comportamento da aprendizagem *online* envolvem a preparação e a MDE diante de arquivos no formato de arquivos de *log*. Os valores em falta na base de dados podem comprometer o resultado a ser obtido nas próximas etapas da abordagem, a exemplo da construção da rede bayesiana. Para tanto, é necessário iniciar a análise a partir da identificação das variáveis do experimento a fim de melhor postular as hipóteses do estudo de caso e construir a árvore de decisão, ilustrado na Tabela 1. Definidas as variáveis do experimento, analisou-se o comportamento da amostra, por meio da análise de algumas características, tais como: distribuição, dispersão e a variância dos dados, o que permitiu observar que os dados não se comportam de maneira linear e não seguem uma distribuição Normal.

Tabela 1: Variáveis identificadas na base de dados.

VARIÁVEL	TIPO VARIÁVEL	DESCRIÇÃO
Nota Final	**Quantitativa**	**Média final do aluno**
Status	**Qualitativa**	**Situação final do aluno após as interações com o AVA.**
Data de inserção	**Quantitativa**	**Data de cadastro do aluno**
Primeiro Acesso	**Quantitativa**	**Data do primeiro acesso**
Acessos	**Quantitativa**	**Quantidade de acessos no semestre do aluno**
Taxa de Participação	**Quantitativa**	**Razão da quantidade de acessos por semana: *T* = Acessos / *(Quantidade de Semanas de Interação do Aluno)***

4.2. Classificação dos Dados Coletados

Seguindo a metodologia proposta, a próxima etapa executada no estudo foi a de classificação dos dados. Esta fase tem o objetivo de esclarecer, diante das variáveis coletadas, qual a hierarquia de importância e qual a variável mais significativa para as saídas esperadas por algoritmos de classificação previamente estabelecidos. Para esta fase, foi utilizado o *software* Weka (árvores de decisão e redes bayesianas). Em seguida, procedeu-se a análise da eficiência dessas técnicas no contexto do trabalho. Alguns critérios foram adotados para comparar tais técnicas, sendo esses: as Instâncias Corretamente Classificadas (ICC) no treinamento, as Instâncias Corretamente Classificadas (ICC) na classificação, as Instâncias Erroneamente Classificadas (IEC) na classificação e estatística *Kappa* calculada.

a) Classificação por Árvores de Decisão

Foram geradas as árvores de decisão utilizando os seguintes algoritmos: *Id3*, *J48*, *BFTree* e *SimpleCart*, com nível de confiança de 99%. Neste trabalho, optou-se pela classificação baseada na validação cruzada (*cross-validation* com 10 *folds*) por permitir

a previsão do comportamento da rede no futuro por meio da avaliação da exatidão da classificação obtida. Os resultados são apresentados na Tabela 2. Ao analisar os resultados da tabela, vale destacar inicialmente que o valor da estatística *Kappa* se mantém em todos os algoritmos, significando que a previsão proveniente da árvore de decisão gerada por esses algoritmos é melhor em 92,73% do que uma previsão randômica.

Tabela 2: Resultados obtidos na geração das árvores de decisão.

TÉCNICA	Id3	J48	BFTree	Simple Cart
ICC no treinamento	97,46%	96,15%	96,15%	96,15%
ICC na classificação	96,15%	96,15%	96,15%	96,15%
IEC na classificação	3,85%	3,85%	3,85%	3,85%
Estatística Kappa na classificação	0.93	0.93	0.93	0.93

Ao analisar a métrica Instâncias Corretamente Classificadas (ICC), percebe-se que o algoritmo Id3, na fase de treinamento, proporcionou resultados melhores do que na fase de classificação, significando que este algoritmo não é a melhor solução para este problema. Os índices de acerto elevados obtidos com os outros algoritmos dão uma maior confiança ao tomador de decisão em analisar a árvore gerada. Outro aspecto que merece destaque é que a árvores geradas pelos métodos J48, *BFTree* e *SimpleCART* no contexto analisado foram bem semelhantes.

b) Classificação por Redes Bayesianas

A classificação, por meio de redes bayesianas, foi levada a efeito a partir do *BayesNet* e o *NaiveBayes*, disponíveis no *Weka* (tabela 3). O classificador *Naive Bayes* é denominado ingênuo por assumir que os atributos são condicionalmente independentes. Outra característica do algoritmo *NaiveBayes* é que este não necessita de parâmetros de entrada. Por isso, testou-se também o *BayesNet*, que constrói uma rede bayesiana completa e realiza a busca de acordo com um algoritmo de busca qualquer. O principal parâmetro para o algoritmo *BayesNet* é o tipo de busca realizado.

Tabela 3: Resultados obtidos na geração da rede bayesiana.

Técnica	*BayesNet*	*NaiveBayes*
ICC no treinamento	97.4359 %	97.4359 %
ICC na classificação	97.4359 %	97.4359 %
IEC na classificação	2.5641 %	2.5641 %
Estatística Kappa na classificação	0.9521	0.9521

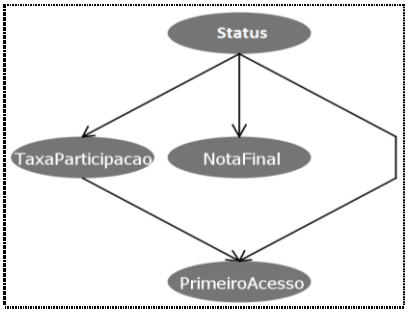

Figura 2. Grafo da rede bayesiana gerada pelo método *BayesNet*.

Diante dos resultados, é possível perceber um comportamento semelhante no desempenho das redes analisadas. Novamente, a estatística *Kappa* se manteve a mesma nas duas redes analisadas, significando que a previsão da rede é melhor em 95,21% do que uma previsão randômica. O índice de acerto nas fases de treinamento e de classificação também é equivalente entre as técnicas analisadas. Outro aspecto que precisa ser comentado é a semelhança entre os grafos gerados pelos métodos. Na Figura 2, é apresentado o grafo gerado pelo algoritmo *NaiveBayes*. O grafo mostra que as variáveis mais relevantes para a saída (variável do *Status*) são a *TaxaParticipação* e *NotaFinal*.

4.3. Interpretação dos Resultados

Diante dos resultados apresentados pelos algoritmos de classificação, a abordagem utilizada mostrou que a aplicação de técnicas de MDE pode enriquecer e auxiliar o professor nas conclusões relacionadas ao ambiente de educação administrado. Nesse sentido, retomam-se as perguntas iniciais: Quais elementos do AVA podem influenciar no bom desempenho do aluno? Foi possível extrair algum comportamento, que não havia sido detectado anteriormente nas interações convencionais com o ambiente?

Inicialmente, o grafo gerado pelas redes bayesianas analisadas ratificou ao professor o entendimento de que, de fato, o *Status* do aluno é determinado por uma série de elementos como: *TaxaParticipação*, *NotaFinal* e *PrimeiroAcesso*. Além disto, o professor percebeu, mediante a análise das árvores de decisão, um comportamento que intuitivamente ele acreditava existir: os alunos que participaram ativamente do ambiente tiveram bons resultados ao final do curso. Isto motivou o professor, parceiro da disciplina, a melhorar as atividades interativas no ambiente e prestar mais atenção nesse quesito com as turmas posteriores. Contudo, o critério Nota ainda é fundamental ao bom desempenho do aluno e as análises demonstram tal fato.

5. Considerações Finais

Tais métodos se mostraram eficientes na análise das informações. A principal contribuição desta abordagem proposta é sua simplicidade de implantação e a inferência gráfica na análise dos resultados, além de auxiliar nos apoio a decisão do professor no período de avaliação do curso EaD lecionado. Além disto, a pesquisa possui algumas ameaças de validade que precisam ser esclarecidas. É possível detectar que há algumas ameaças à validade no experimento: a falta de randomização dos sujeitos visto que o experimento é um estudo de caso, baixo poder estatístico devido às amostras serem pequenas e à limitação de se trabalhar apenas com uma plataforma de EaD (*Moodle*).

O próximo passo deste trabalho consiste em aplicar outras técnicas de aprendizado para avaliar se as informações obtidas serão ratificadas. Outras análises podem advir desse estudo inicial, tais como: a análise detalhada sobre os tipos e a qualidade da interação no ambiente. Outra possível área de atuação visa a incorporar essa análise de *Status* aos AVA de maneira geral para que o professor acompanhe a taxa de participação do aluno, pois atualmente somente por meio de relatórios o docente observa a frequência do aluno no ambiente.

Este trabalho serve, portanto, como um passo inicial e importante ao um projeto mais amplo voltado ao auxílio na tomada de decisão de um professor (ou tutor) diante de contextos tão diferenciados como os presentes em AVA.

Referências

Amorim, M.T. ; Cury, D. ; Menezes, C. S. (2011). Um sistema inteligente baseado em ontologia para apoio ao esclarecimento de dúvidas. In: XXII Simpósio Brasileiro de Informática na Educação, 2011, Aracaju. Anais do SBIE-2011.

Baker, R.S.J. (2011) Data Mining for Education. International Encyclopedia of Education, 3rd ed., edited by B. McGaw, P. Peterson, and E. Baker. Oxford, UK: Elsevier.

Chikalov, I. (2011). Average Time Complexity of Decision Trees. Springer. Disponivel em:<http://www.springerlink.com/content/978-3-642-22661-8#section=938076&page=1>. ISBN 978-3-642-22660-1.

Corrigan, J. A. (2012). The implementation of e-tutoring in secondary schools: A diffusion study. Computers & Education. 59 (3) pages 925–936. Elsevier.

Russell, S., Norvig, P. (2004) Artificial Intelligence – A Modern Approach, Prentice-Hall, 2a Edição.