Uma Abordagem Genérica de Identificação Precoce de Estudantes com Risco de Evasão em um AVA utilizando Técnicas de Mineração de Dados

Ramon Nóbrega dos Santos
Universidade Federal da Paraíba
João Pessoa - Brasil
ramonob13@gmail.com

Clauirton de Alburqueque Siebra
Universidade Federal da Paraíba
João Pessoa - Brasil
clauirton@di.ufpb.br

Estêvão Domingos Soares Oliveira
Universidade Federal da Paraíba
João Pessoa - Brasil
estevaodso@gmail.com

ABSTRACT
This paper describes a generic approach in order to early identify students at risk of dropout in a Learning Management System (LMS) using data mining techniques. The advantage of this approach is the use of only time-varying data, thus avoiding the use of questionnaires to collect data. The proposal is presented as an architecture to predict student dropout in distance graduation courses. The experiments performed via decision tree algorithms provided an average accuracy of 81,55% using the first semester grades.

Keywords
Dropout, Ambiente Virtual de Aprendizagem, Data Mining, Árvores de Decisão, Moodle

RESUMO
Este trabalho descreve uma abordagem genérica com o objetivo de identificar precocemente estudantes com risco de evasão em um Ambiente Virtual de Aprendizagem (AVA) utilizando técnicas de mineração de dados. A proposta é apresentada como uma arquitetura para predizer a evasão de estudantes em cursos de graduação à distância. A vantagem desta abordagem é a utilização de apenas dados variantes no tempo, evitando desta forma a utilização de questionários para a coleta de dados. Os experimentos realizados a partir de algoritmos de árvores de decisão forneceram precisões médias de 81,55% utilizando as primeiras notas semestrais.

Palavras-chave
Evasão, Ambiente Virtual de Aprendizagem, Mineração de Dados, Moodle.

1. INTRODUÇÃO
De forma a diminuir a evasão de estudantes, alguns trabalhos se direcionaram a entender as razões que levam a esta evasão, identificando os estudantes que tendem a apresentar este comportamento. Pesquisas iniciais utilizaram estudos qualitativos, comportamentais e baseados em questionários. Esses estudos desenvolveram diversas teorias para este fenômeno, entretanto, não foi proposto nenhum instrumento para precisamente predizer, e potencialmente diminuir, à evasão dos estudantes [11,10]. Dessa forma, uma nova linha de estudo, baseada em técnicas de mineração de dados, vem sendo utilizada na identificação de estudantes propensos à evasão.

A motivação do presente trabalho é desenvolver uma abordagem genérica de identificação de estudantes com risco de evasão em um AVA que possa ser aplicada nos mais diferentes contextos uma vez que utilizará dados que todos os cursos possuem: as notas parciais das disciplinas de um AVA e as notas finais das disciplinas aos finais dos períodos.

O objetivo da predição é estimar um valor desconhecido de uma variável que descreve o estudante. Na educação, os valores normalmente preditos são desempenho, conhecimento, pontos ou notas. Este valor pode ser numérico ou contínuo (tarefa de regressão) e categórico/discreto (tarefa de classificação). A predição de desempenho de um estudante é uma das mais antigas e populares aplicações da Mineração de Dados (*Data Mining*) na educação e diferentes técnicas e modelos têm sido usados (redes neurais, redes Bayesianas, sistemas baseados em regras, regressão e análise de correlação).

Normalmente, a predição de desempenho é realizada para: descobrir grupos potenciais de estudantes com características similares e reações a uma particular estratégia pedagógica [4], para detectar maus usos dos sistemas [1], encontrar grupos de estudantes com determinado comportamento e encontrar equívocos comuns que os estudantes cometem [12], identificar estudantes com baixa motivação e desenvolver ações preventivas para evitar a evasão discente [5]. A utilização de algoritmos de mineração de dados em dados educacionais para a previsão da situação acadêmica é um campo de investigação ainda não consolidado, o qual necessita de investigações complementares tanto na definição dos atributos a serem utilizados quanto nas técnicas de mineração de dados empregados [3].

O problema da predição de evasão de um estudante pode ser analisado como um problema de predição de desempenho no qual são consideradas duas classes principais: evadido ou graduado. Neste trabalho, a classe "evadido" refere-se a um estudante que não conseguiu concluir o curso seja por um abandono, por insuficiência no desempenho acadêmico ou uma solicitação formal. Já a classe "graduado" refere-se a um estudante que concluiu todas as etapas para o término do curso.

Existem diversos estudos na educação a distância que focam na predição de desempenho de alunos utilizando logs obtidos de Ambientes Virtuais de Aprendizagem (AVAs). Os trabalhos que realizam predição de desempenho podem ser divididos em dois grupos principais: os que utilizam dados invariantes no tempo e/ou dados variantes no tempo. Os dados invariantes no tempo são os que não podem ser modificados no tempo, já os dados variantes no tempo são os que podem ser modificados. Normalmente, os dados invariantes são dados socioeconômicos, demográficos e obtidos a partir de questionários. Os trabalhos indicam que modelos preditivos que utilizam dados invariantes no tempo

fornecem precisões inferiores quando comparados com os modelos que utilizam dados variantes no tempo [8].

Normalmente, os dados variantes no tempo são os que podem ser obtidos a partir do monitoramento do estudante na plataforma educacional de um AVA. Uma das principais vantagens da utilização de apenas dados variantes na construção dos modelos preditivos é a não necessidade da utilização de questionários para a obtenção de dados dos alunos. Dessa forma, torna-se menos trabalhosa a etapa de pré-processamento e a etapa de transformação dos dados. A outra grande vantagem é a possibilidade de aplicar a mesma abordagem nos mais diferentes tipos de cursos a distância, tornando a abordagem genérica.

No trabalho [8] é proposto um método de predição de desempenho em cursos a distância utilizando redes neurais. Uma das diferenças do trabalho de Lykourntzoul et al (2009) para a presente abordagem é que em Lykourntzoul et al (2009) é focado na previsão da desempenho de apenas uma disciplina de um curso a distância. Já no presente trabalho é considerada a predição de evasão de um curso de graduação a distância de maior duração. Outra diferença é que em [8] são utilizados dados invariantes e variantes no tempo, já no presente trabalho são propostos somente o uso de dados variantes.

Outro trabalho que desenvolve um método de predição de desempenho em um curso a distância é o [7] onde são usados dados invariantes e variantes no tempo. Primeiramente são utilizados os dados invariantes no tempo e com o passar do tempo vão sendo incorporados dados variantes. Nos primeiros momentos são utilizados dados sócio-demográficos e no decorrer do tempo são incorporados dados de logs. No contexto do presente trabalho de propor uma arquitetura para predição de aluno com risco de evasão em um curso a distância de graduação, a principal diferença do trabalho [7] é a utilização restrita a predição de desempenho de apenas uma disciplina.

Nos trabalhos mencionados [7,8] são focados na previsão de desempenho de apenas uma disciplina, o que não possibilita fazer uma associação direta entre o insucesso nela e no curso de graduação, ou seja, nem sempre um aluno que reprova determinada disciplina, evadirá do curso como um todo. Esses dois estudos [7,8] consideram que apenas o desempenho do estudante ao final da disciplina sem levar em consideração a variável preditiva ao final do curso: a graduação ou evasão do aluno. Outro ponto a ser destacado é que nenhum dos dois trabalhos [7,8] propôs a utilização de apenas dados variantes no tempo para a construção de modelos preditivos de identificação de um aluno com risco de evasão em um curso a distância de maior duração. Nesses estudos são mostrados que a utilização de dados invariantes no tempo não traz precisões melhores do que com a utilização de dados variantes no tempo.

Portanto, é interessante investigar a utilização de uma abordagem genérica de identificação de um aluno com risco de evasão utilizando somente dados variantes no tempo que possui a facilidade na etapa de pré-processamento, pois não necessita da utilização de questionários e tem o caráter genérico, pois esses modelos poderão ser usados em diferentes contextos.

2. ABORDAGEM GENÉRICA DE IDENTIFICAÇÃO DE EVASÃO

A Abordagem Genérica de Identificação de Evasão proposta foca na utilização de apenas dados variantes no tempo, pois pretende ser genérica para ser aplicada nos mais diferentes contextos. A abordagem genérica de identificação de estudantes a distância em um Ambiente Virtual de Aprendizagem (AVA) com risco de evasão utiliza o monitoramento de dados de um AVA com a utilização de notas intermediárias das atividades das disciplinas do período ou semestre letivo para prever o desempenho final do aluno na disciplina considerada, aqui denominados de "*Modelos na*". Quando as médias finais das disciplinas já estão disponíveis no banco de dados do Sistema de Controle Acadêmico (SCA), são utilizados modelos que fazem a previsão de um aluno com risco de evasão no curso de graduação utilizando as notas aos finais dos períodos, aqui denominados de "*Modelos nb*". Na Figura 1 é vista a arquitetura da abordagem proposta.

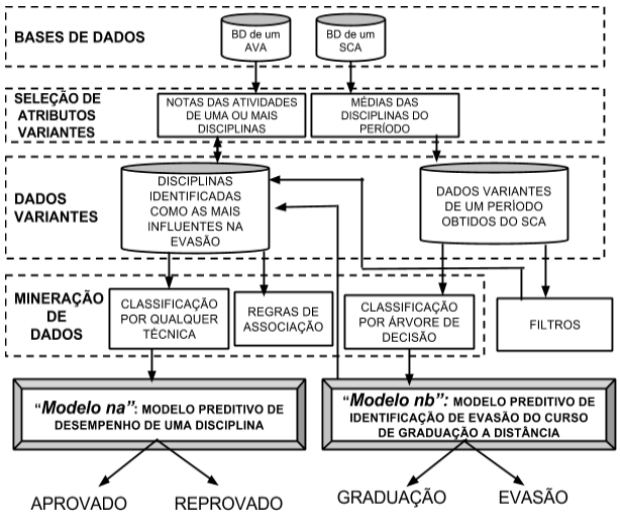

Figura 1. Arquitetura da Abordagem Genérica de Predição de Evasão Proposta

As bases de dados propostas nos modelos preditivos são duas: a base de dados de um AVA e a base de dados de um SCA. A fase de Pré-processamento é realizada a partir da *seleção de atributos variantes* composta por duas etapas: seleção das notas das atividades da disciplina escolhida para previsão do desempenho e seleção das médias e situações das disciplinas de determinado período para a predição da evasão do curso de graduação a distância. Com os dados pré-processados, poderão ser utilizados algoritmos de mineração de dados por Árvores de Decisão com o objetivo da criação de um modelo preditivo que identificará precocemente se um aluno será evadido ou graduado ou também para identificação dos atributos mais importantes relacionados com a evasão.

A partir da utilização de filtros também será possível à descoberta dos atributos com maior peso na evasão. Possivelmente serão identificadas as disciplinas que mais influenciam na evasão, dessa forma, essas disciplinas podem ser analisadas na predição dos desempenhos dos alunos nas mesmas. Os algoritmos de Regras de Associação serão investigados a fim de encontrar regras úteis baseadas nas atividades dos estudantes que possam auxiliar o professor nas melhoras dos desempenhos dos alunos. Na arquitetura são usados os algoritmos de Classificação onde são utilizadas técnicas computacionais para separar os dados em uma base de dados de treinamento e uma base de dados de teste.

A técnica de estratificação de dados que foi usada neste trabalho é a Validação Cruzada de Dez Partições em que os dados são divididos em dez partições aleatórias sendo retiradas nove dessas partições para serem utilizados no conjunto de treinamento e uma partição para o conjunto de testes. Dessa primeira iteração é obtida a primeira precisão do modelo. Depois são realizadas, para

cada algoritmo de classificação aplicado, mais nove iterações, percorrendo todas as possibilidades de escolha, resultando em mais nove valores de precisão. A precisão final do Classificador é calculada, considerando a média das precisões das dez iterações. Ao modelo gerado pelo Classificador daremos o nome de modelo preditivo, pois ele irá predizer se um estudante evadiu ou graduou-se considerando o curso completo de graduação a distância ou se foi aprovado ou reprovado considerando uma disciplina.

A Abordagem Genérica de Identificação de Estudante com Risco de Evasão divide-se em duas etapas:

1) Caso o aluno ainda não tenha a média da disciplina, esta poderá ser inferida a partir das suas notas parciais ao longo das atividades do curso de um AVA (*Modelos na*). Com a inferência das notas de todas as disciplinas, será possível inferir a média final do período. Dessa forma, os modelos preditivos poderão ser aplicados já em momentos iniciais do curso.

2) Com o período finalizado, poderão ser utilizadas as médias finais das diferentes disciplinas que compõe o período para predizer a evasão ou graduação do aluno (*Modelos nb*).

A etapa 1 refere-se a utilização de dados de um AVA, analisada na seção 2.1. Já na etapa 2 são utilizados dados de um SCA, vista na seção 2.2.

2.1 Abordagem Genérica a partir de Dados de um AVA (Modelos na)

A etapa 1 corresponde ao processamento de logs de um AVA, que no presente estudo utiliza o Moodle. São propostas as seguintes tabelas do Moodle, conforme mostrado abaixo (Tabela 1).

Tabela 1. Tabelas utilizadas do AVA do Moodle.

Tabela	Descrição
mdl_user_students	Informações sobre todos os estudantes
mdl_log	Informações sobre logs dos estudantes
mdl_grades	Informações sobre as notas dos estudantes

A partir de consultas às tabelas *mdl_user_students*, *mdl_log* e *mdl_grades* serão obtidas as notas parciais dos estudantes que na presente abordagem são agrupadas por atividades. Cada atividade corresponde a uma tarefa realizada pelo aluno que possui uma nota. Na Tabela 2 é vista como as notas serão organizadas para que seja possível a predição do desempenho do estudante e consequentemente a previsão de evasão do curso como um todo.

Tabela 2. Modelo proposto para prever desempenho do estudante a partir de um AVA no "*Modelo na*"

Atividade	1	2	3	m	Situação final da disciplina w
Notas do aluno 1	9	7	5	10	Aprovado
Notas do aluno 2	2	3	8	6	Reprovado
Notas do aluno 3	1	3	7	4	Reprovado
Notas do aluno n	7	5	9	7	Aprovado

Cada linha da Tabela 2 apresenta as notas das atividades (1, 2, 3 e *m*) do aluno 1 ao aluno *n*. A partir da utilização de algoritmos de Classificação com as notas das atividades iniciais é possível prever antecipadamente se um aluno será aprovado ou reprovado na disciplina. A partir da utilização de todas as disciplinas de um período será possível prever a média final do período. Dessa forma, de forma indireta, poderá ser prevista a evasão de um aluno no curso de graduação a distância a partir de previsões de desempenhos de disciplinas. No momento atual do trabalho estão sendo selecionadas bases de dados para a realização dos experimentos e sendo realizadas atividades de extração de dados do AVA. No exemplo hipotético da Tabela 2, são realizadas as etapas de transformação da tabela em um arquivo arff (Quadro 1) e aplicação dos algoritmos de mineração de dados pela técnica de Classificação.

Quadro 1 – Arquivo arff da Tabela 2

```
@relation notas_atividades
@attribute nota_atividade1 numeric
@attribute nota_atividade2 numeric
@attribute nota_atividade3 numeric
@attribute nota_atividaden numeric
@attribute desempenho {APROVADO,REPROVADO}
```

Foi utilizado neste exemplo hipotético da Tabela 2, na qual são mostradas as notas das atividades de uma disciplina hipotética, o algoritmo de Árvore de Decisão J48 (um dos três utilizados nos experimentos deste trabalho), pois a partir dele é possível visualizar a árvore de decisão construída, conforme mostrado no Quadro 2.

Quadro 2 – Árvore de Decisão obtida com a aplicação do algoritmo J48 a partir do arquivo .arff da Tabela 2

```
nota_atividade1 <= 2: REPROVADO (2.0)
nota_atividade1 > 2: APROVADO (2.0)
```

A leitura do Quadro 2 é realizada da seguinte forma: alunos que tiraram nota_atividade1 (nota na atividade 1) menor ou igual a dois foram reprovados ao final da disciplina e os demais alunos foram aprovados. A partir da aplicação de modelos preditivos para todas as disciplinas será possível prever a classe final de média do aluno. Na presente abordagem serão utilizadas quatro classes observando os desempenhos dos alunos: nota<5: baixo; 5<nota<7: regular; 7<nota<8,5: ótimo; nota>=8,5: excelente. A partir da utilização de mais classes do que as mostrados no exemplo hipotético do Quadro 1, que foram duas "aprovado" e "reprovado", será possível prever com maior exatidão a média final, ao final do período, a partir de todas as disciplinas.

2.2 Abordagem Genérica a partir de Dados de um SCA (Modelos nb)

Na Tabela 3 são mostradas as variáveis propostas utilizadas aos finais dos períodos. O atributo "situação da disciplina" pode assumir quatro valores: aprovado, reprovado por nota, reprovado por falta ou indefinido. A disciplina é considerada "aprovada" quando o aluno obtém média final na disciplina maior ou igual a cinco. A disciplina é considerada "reprovada" quando o aluno obtém média final menor do que cinco. A disciplina é considerada "reprovada por falta" quando o aluno não realiza as provas da disciplina. A disciplina é considerada "indefinida" quando o aluno não cursou a disciplina.

Tabela 3. Variáveis utilizadas ao final do período obtidas do SCA

Variável	Valores
Situação da disciplina	(aprovado, reprovado por nota, reprovado por falta ou indefinido)
Nota da disciplina	Entre 0 e 10
Quantidade de reprovações no período ou semestre	Entre 0 e a quantidade de disciplinas matriculadas no semestre
Média no período	Entre 0 e 10

A base de dados foi composta pelas notas das disciplinas e a situação final (aprovado, reprovado por nota e reprovado por falta em cada disciplina), a média no período e, por fim, o atributo identificador da classe de aluno: graduado ou evadido, considerando o curso de graduação. Na Figura 2 pode ser observada a abordagem genérica proposta. As informações referem-se às entradas necessárias para a construção do modelo preditivo. Com o modelo preditivo construído, poderão ser antecipadamente identificados alunos com risco de evasão.

O *Modelo 1a* refere-se ao modelo preditivo que prevê o desempenho de um aluno em uma disciplina do primeiro período. O *Modelo 1b* refere-se ao modelo preditivo que prevê a evasão de um aluno ao final do primeiro período. O *Modelo 2a* refere-se ao modelo preditivo que prevê o desempenho de um aluno em uma disciplina do segundo período. O *Modelo 2b* refere-se ao modelo preditivo que prevê a evasão do aluno ao final do segundo período. O *Modelo na* refere-se ao modelo preditivo que prevê o desempenho de um aluno em uma disciplina do período n. O *Modelo nb* refere-se ao modelo preditivo que prevê a evasão de um aluno ao final do período n. A vantagem da utilização dos modelos intermediários (*Modelos na*) é que não é necessário esperar pela implantação das notas finais das disciplinas em determinado período, uma vez que ações já poderão ser tomadas para evitar reprovações dos alunos nas disciplinas analisadas. Assim, a previsão da evasão ou graduação de um curso à distância do aluno poderá ser antecipada.

Na Figura 2 é mostrada a Arquitetura Temporal da Abordagem Preditiva proposta. Denominamos de arquitetura temporal, pois ela utiliza de dados variantes no tempo com a aplicação de modelos preditivos em diferentes instantes de tempo do curso de graduação. A partir de dados de treinamento é possível a construção dos modelos preditivos. E quando os mesmos estiverem construídos, será possível prever a aprovação ou reprovação em uma disciplina e a evasão ou graduação no curso de graduação a distância.

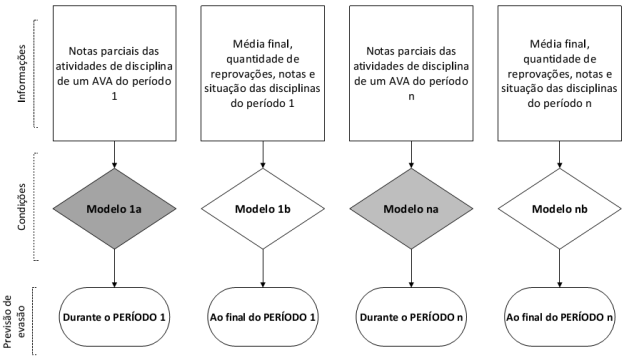

Figura 2 - Arquitetura Temporal da Abordagem Preditiva de Evasão Proposta

A característica principal da presente arquitetura temporal é possibilitar que outros períodos possam fornecer informações importantes sobre o risco de evasão de um aluno, a partir da identificação das disciplinas que mais influenciam na evasão em diferentes períodos. Assim, o presente trabalho propõe, de forma inovadora, uma abordagem temporal para identificar um aluno com risco de evasão em diferentes momentos de um curso de graduação a distância utilizando um AVA.

3. METODOLOGIA DE APLICAÇÃO DOS MODELOS PREDITIVOS

A construção dos modelos preditivos foi realizada a partir da ferramenta Weka. Cada algoritmo é executado 10 vezes e seu desempenho final é obtido a partir da média das execuções. No caso do 10-fold cross validation significa que um classificador foi executado 10 vezes para os conjuntos de treinamento e de teste.

Foi usado o ambiente da ferramenta Weka: o Weka Explorer (WE). Este ambiente permite a seleção e execução de um algoritmo classificador por vez. Sendo assim, o resultado do experimento representa a acurácia de uma rodada do algoritmo onde foram realizadas comparações nas acurácias dos classificadores. O WE oferece quatro opções de estratificação da base de dados: *use training set*, *supplied test set*, *cross-validation* e *percentage split*. Foi escolhida a estratificação por cross-validation.

A escolha dos algoritmos pelas técnicas de Árvore de Decisão surgiu da necessidade de descobrir a influência que cada atributo possui no resultado final da classe, sendo os mesmos considerados de Caixa Branca. Algoritmos de Classificação por Caixa Branca obtém modelos que podem explicar as predições por regras do tipo: SE-ENTÃO, permitindo a explicação do funcionamento interno do modelo. As regras do tipo "SE-ENTÃO" são uma das formas mais populares de representação do conhecimento, por causa da sua simplicidade e compreensibilidade. Esses tipos de regras são facilmente entendidas e interpretadas por atores da educação (gestores, professores, tutores, etc) nos mais diferentes propósitos.

4. EXPERIMENTOS REALIZADOS E RESULTADOS

Os experimentos foram realizados a partir da retirada de dados dos alunos do Sistema de Controle Acadêmico (SCA) da Unidade de Educação à Distância da Universidade Federal da Paraíba, também conhecida como UFPB Virtual, que integra o sistema de Universidade Aberta do Brasil (UAB). Foram considerados todos os alunos do curso de Letras da UFPB Virtual que ingressaram nos períodos de 2007.2, 2008.1 e 2008.2 totalizando 1046 alunos: 464 que se graduaram e 562 que evadiram. Foram considerados os períodos de 2007.2 a 2012.2 que consiste em um intervalo de doze períodos (tempo suficiente para observar a evasão ou graduação de um aluno). O tempo normal de conclusão do curso de Letras da UFPB Virtual é de oito períodos.

A base de dados foi dividida em duas classes distintas. A primeira classe é composta por alunos que completaram todos os requisitos para a conclusão do curso, os graduados. A segunda classe é composta por alunos que não concluíram o curso por iniciativa própria (abandono ou trancamento de matrícula); ou por imposição da universidade (reprovação por nota, ultrapassar o prazo de conclusão do curso ou solicitação formal), os evadidos.

Com os dados desses alunos obtidos do SCA, foram realizadas atividades de pré-processamento e transformação dos dados para deixá-los no formato arff proposto (conforme variáveis mostradas

na Tabela 3) que é um formato de arquivo apropriado para realização de mineração de dados no ambiente Weka. No arquivo arff, cada linha corresponde a uma instância de um aluno. Na Tabela 4 são mostradas as disciplinas que compõe o primeiro período do curso de Letras da UFPB Virtual.

Tabela 4. Lista de disciplinas do primeiro período do curso de Letras da UFPB Virtual

Id da Disciplina	Nome da disciplina
1	FUND ANTROPO-FILOS.EDUCACO - UV
2	INT. AOS ESTUDOS LITERARIOS - UV
3	INT. AOS ESTUDOS CLASSICOS - UV
4	LEITURA E PRODUCAO DE TEXTOS I - UV
5	INTRODUCAO A EDUCACAO A DISTANCIA – UV
6	FUNDAMENTOS DE LINGUISTICA - UV

4.1 Acurácia Geral

O critério que foi utilizado neste estudo para medir as precisões das predições obtidas pelos classificadores é o da acurácia. O critério da acurácia geral é usado para medir a proporção total dos estudantes com situação final, evadido ou graduado, que foi corretamente predita pela técnica. O critério é usado para medir o número de estudantes corretamente classificados da classe de graduados mais o número de estudantes corretamente classificados da classe de evadidos, dividido pelo número total de estudantes.

4.2 Experimentos com todos os atributos

Os atributos considerados foram referentes ao primeiro período, sendo os seguintes: média do primeiro período, quantidade de disciplinas reprovadas no primeiro período, média da disciplina 1 indo até a média da disciplina 6, situação da disciplina 1 indo até a situação da disciplina 6. Para cada aluno foram recolhidas todas as suas disciplinas cursadas no primeiro período, as médias e situações dessas disciplinas e a quantidade de disciplinas reprovadas. Caso o aluno não tenha cursado alguma das disciplinas do primeiro período, foi atribuído o valor 0 para a disciplina e a situação "indefinido" para essa disciplina.

A base de dados foi dividida em dez conjuntos utilizando o método da validação cruzada (10 fold cross-validation). Os algoritmos, aplicados a base de dados, foram executados dez vezes, valor padrão de configuração do ambiente. Foram aplicados os seguintes algoritmos de classificação por árvore de decisão: o SimpleCart (SC), o J48 (J48) e o ADTree (AT). A ferramenta Weka calculou as médias das acurácias obtidas em cada rodada dos classificadores conforme mostradas na Tabela 5.

Tabela 5. Acurácia e taxas dos classificadores considerando todos os atributos

Classificador	SC		J48		AT	
Acurácia	81,45%		82,60%		81,83%	
Matriz de Confusão	447	135	484	98	456	126
	59	405	84	380	64	400

Na Tabela 5 são mostradas as acurácias dos classificadores para o conjunto de teste e a matriz de confusão, composta pela classe positiva, alunos que concluíram o curso (graduados), e negativa, alunos que não concluíram o curso (evadidos).

A acurácia média dos três classificadores foi de 81,55%. A leitura da matriz de confusão é realizada da seguinte forma: considerando o exemplo da matriz de confusão gerada pelo algoritmo de Classificação SimpleCart: 447 alunos foram corretamente classificados como evadidos, 405 alunos foram corretamente classificados como graduados, 135 alunos foram incorretamente classificados como graduados (de fato, evadiram do curso) e 59 alunos foram incorretamente classificados como evadidos (de fato, se graduaram). Essa mesma leitura pode ser realizada para as demais matrizes de confusão deste trabalho.

Foram geradas árvores de decisão com tamanhos altos, a árvore de decisão gerada pelo algoritmo J48 teve 68 nodos. O algoritmo ADTree gerou uma árvore com 31 nodos. Já o algoritmo SimpleCart gerou uma árvore menor com apenas 6 nodos, conforme mostrado no Quadro 3. A árvore de decisão pode ser lida da seguinte forma: um aluno que tirou média final no primeiro período menor do que 5,36, evadiu do curso (444 alunos se enquadram nesta regra e 39 alunos não se enquadram). Já para os alunos que tiraram média no primeiro período maior do que 5,36, outros atributos influenciaram na graduação ou evasão do aluno.

Quadro 3 – Árvore de Decisão obtida com a aplicação do algoritmo SimpleCart considerando todos os atributos

```
CART Decision Tree

media_final1 < 536.0: ABANDONO(444.0/39.0)
media_final1 >= 536.0
|  nota5 < 675.0
|  |  nota4 < 716.5
|  |  |  nota6 < 853.0
|  |  |  |  nota1 < 725.0: GRADUACAO(22.0/17.0)
|  |  |  |  nota1 >= 725.0: ABANDONO(31.0/10.0)
|  |  |  nota6 >= 853.0: GRADUACAO(11.0/2.0)
|  |  nota4 >= 716.5: GRADUACAO(12.0/1.0)
|  nota5 >= 675.0: GRADUACAO(370.0/87.0)

Number of Leaf Nodes: 6

Size of the Tree: 11
```

Diante das árvores com muitos nodos resultantes dos algoritmos J48 e ADTree, surgiu a necessidade de utilizar filtros para a escolha dos atributos mais relevantes para encontrar árvores de decisões com quantidade de nodos menores, o que possibilita a geração de regras mais simples.

4.3 Experimentos usando filtros para escolha dos atributos mais relevantes

Nos experimentos desta seção foram aplicados filtros que tentam identificar quais atributos têm mais impacto na variável preditiva final (evadido ou graduado). Os filtros têm a característica de avaliar os atributos independentemente do algoritmo de aprendizagem. O Weka fornece vários filtros, dos quais, foram escolhidos os seguintes: SymmetricalUncertAttributeEval, CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttributeEval, FilteredSubsetEval e InfoGainAttributeEval. Todos os filtros identificaram o atributo media_periodo1 como o mais importante. Os demais atributos escolhidos com melhores pontuações foram os seguintes, em ordem crescente: quantidade de disciplinas reprovadas no primeiro período, média da disciplina 5 e média da disciplina 6. Ou seja, as disciplinas INTRODUCAO A EDUCACAO A DISTANCIA e FUNDAMENTOS DA LINGUISTICA foram as identificadas como as que mais influenciam na evasão ou graduação do curso a distância como um todo. Assim, foram executados os seguintes algoritmos de Classificação por árvore de decisão: J48, SimpleCart e ADTree considerando somente os atributos escolhidos pelos filtros. A

ferramenta calculou as médias das acurácias obtidas em cada rodada dos classificadores mostradas na Tabela 6.

Tabela 6. Acurácia e taxas dos classificadores considerando os atributos selecionados pelos filtros

Classificador	SC		J48		AT	
Acurácia	81,64%		80,68%		80,49%	
Matriz de Confusão	453	129	459	123	429	153
	63	401	79	385	51	413

Dos resultados obtidos, vale destacar que houve uma melhora de acurácia no algoritmo SimpleCart de 81,45% para 81,64%, mesmo com a utilização de apenas um atributo (a média final do primeiro período) onde foi gerada a árvore de decisão do Quadro 4.

Quadro 4 – Árvore de Decisão obtida com a aplicação do algoritmo SC usando os atributos relevantes

```
media_final1 < 536: ABANDONO (444.0/39.0)
media_final1 >= 536.0: GRADUACAO (425.0/138.0)
```

4.4 Discussão dos Resultados

Nesta seção, são discutidos os resultados principais do estudo. Foram excluídos os atributos de tempo invariantes das bases de dados. Analisando os experimentos realizados nas seções 4.2 e 4.3, a acurácia média dos três classificadores utilizados nos experimentos realizados com todos os atributos (seção 4.2) foi de 81,55%. Já com a utilização dos atributos identificados pelos filtros como os mais importantes, a acurácia média dos três classificadores foi de 80,93% (seção 4.3).

Diante dos resultados, podemos concluir que é interessante a aplicação dos filtros, mesmo que possa ocorrer uma diminuição da precisão, pois a vantagem que é proporcionada de terem-se regras mais concisas possibilita a melhor tomada de decisões por parte de gestores e professores. A partir de resultados mais simples, é possível visualizar com mais clareza quais os atributos que mais influenciam na evasão, conforme visto no quadro 4 da seção 4.3.

5. CONCLUSÕES E TRABALHOS FUTUROS

A partir dos experimentos realizados foi constatado que já ao final do primeiro período é possível prever o risco de um aluno evadir com precisão média maior do que 80%. Como trabalhos futuros serão realizados experimentos com o *Modelo 1a* proposto e com os demais *Modelos na* e *Modelos nb*. Serão realizados outros experimentos acrescentando novos algoritmos de Árvores de Decisão. Serão também testados algoritmos de Regras de Indução, pois permitem a geração automática de regras. Também será aplicada a abordagem aqui proposta para outros cursos, além do de Letras da UFPB Virtual utilizado nos experimentos, para verificar se os resultados se repetem.

A maior contribuição deste trabalho é propor uma arquitetura genérica de predição de evasão de um curso a distância de maior duração, que considera todas as disciplinas de um curso à distância, a partir da utilização de apenas dados variantes no tempo em uma abordagem genérica. O presente trabalho também mostrou que com poucos atributos pode-se fazer a previsão de alunos com risco de evasão. As investigações iniciais abrem a possibilidade da utilização de outras técnicas de mineração de dados.

Os benefícios da aplicação de mineração de dados no problema da evasão são os seguintes: i) identificar nos primeiros momentos do curso os alunos mais propensos à evasão; ii) permitir que a universidade não utilize somente análises estatísticas no tratamento do problema da evasão, e; iii) identificar as disciplinas mais associadas com à evasão, permitindo a aplicação de modelos preditivos nos AVAs. A aplicação da abordagem proposta pode também ser utilizada na educação tradicional, a qual não foi examinada no presente estudo.

6. REFERÊNCIAS

[1] BAKER, R. S.; CORBETT, A. T.; KOEDINGER, K. R. **Detecting Student Misuse of Intelligent Tutoring Systems**. Proceedings of the 7th International Conference on Intelligent Tutoring Systems. [S.l.]: Springer Verlag. 2004. p. 531-540.

[2] BAKER, R..; ISOTANI, S.; CARVALHO, A.**Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, Vol. 19, No. 2. p. 2-13, 2011.

[3] CASTRO, F. et al. (2007). **Applying Data Mining Techniques to e-Learning Problems, Studies in Computational Intelligence** (SCI). 62, 183 - 221 (2007) Springer-VerlagBerlinHeidelberg.

[4] CHIEN, C.-F.; CHEN, L.-F. **Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry**.Expert Syst. Appl., v. 34, n. 1, p. 280-290, 2008.

[5] COCEA, M.; WEIBELZAHL, S. **Can Log Files Analysis Estimate Learners` Level of Motivation?**. LWA. [S.l.]: University of Hildesheim, Institute of Computer Science. 2006. p. 32-35.

[6] DEKKER, G., Pechenizkiy M. and Vleeshouwers J. (2009). **Predicting Students Drop Out: A Case Study**. In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, Pages 41-50.

[7] KOTSIANTIS, S. B.; PATRIARCHEAS, K.; XENOS, M. N. **A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education**. Knowl.-Based Syst., v. 23, n. 6, p. 529-535, 2010.

[8] LYKOURENTZOU, I. et al. **Dropout prediction in e-learning courses through the combination of machine learning techniques**. Computers & Education, v. 53, n. 3, p. 950-965, 2009.

[9] MANHÃES, L. et al. **Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados**. Anais do XXII SBIE, Aracaju, 2011.

[10] MANNAN,M,A et al. **Student attrition and academic and social integration: application of Tinto's model at the university of Papua New Guinea**. Higher Education 53 (2) (2007) 147-165.

[11] VEENSTRA.C.P et al. **A strategy for improving freshman college retention** , Journal for Quality and Participation 31 (4) (2009) 19-23.

[12] YUDELSON, M *et al*. **Mining Student Learning Data to Develop High Level Pedagogic Strategy in a Medical ITS**. AAAI Workshop on Educational Data Mining, 2006. pp.1-8.