Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão

Ramon Nóbrega dos Santos[1], Clauirton de Alburqueque Siebra[1], Estêvão Domingos Soares Oliveira[1]

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

ramonob13@gmail.com, clauirton@di.ufpb.br, estevaodso@gmail.com

Abstract. *This paper describes an approach in order to early identify graduate students at risk of dropout in a Learning Management System (LMS) using the data mining techniques of Decision Trees. These techniques allow to understand the internal operation of the model, enabling to discover disciplines that have more influence on dropout of graduation courses. This knowledge can be used to predict the academic performance of these disciplines using the LMS. The experiments performed via decision tree algorithms in Academic Control System (ACS) and LMS provided an average accuracy of 80% using the first semester grades. The experiments performed in a Moodle LMS provided, as best result, an accuracy of 89,47% in predicting performance in a discipline.*

Resumo. *Este trabalho descreve uma abordagem que objetiva a identificação precoce de estudantes de graduação a distância com risco de evasão utilizando técnicas de mineração de dados conhecidas como Árvores de Decisão. Essas técnicas permitem o entendimento do funcionamento interno do modelo, possibilitando a descoberta das disciplinas que mais influenciam na evasão do curso de graduação. Este conhecimento pode ser usado para prever o desempenho acadêmico nestas disciplinas no AVA. Foram utilizados algoritmos de Árvores de Decisão nos experimentos no SCA e o AVA, os quais forneceram acurácias médias de 80% na predição da evasão ou graduação do curso, utilizando apenas as primeiras notas semestrais. No AVA, como melhor resultado, foi obtida uma acurácia de 89,47% na predição do desempenho de uma disciplina.*

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, a evasão dos estudantes (VEENSTRA, 2009; MANNAN, 2007). 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 propor uma

abordagem de identificação de estudantes com risco de evasão em um curso de graduação a distância de maior duração 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 ao longo das semanas e as notas finais das disciplinas ao final dos períodos de um SCA.

A predição de desempenho de um estudante é usada para desenvolver ações preventivas para evitar a evasão discente (COCEA, 2006). Os trabalhos que realizam predição de desempenho e/ou evasão 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. 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 trazem precisões inferiores quando comparados com os modelos que utilizam dados variantes no tempo (LYKOURENTZOU, 2008). Normalmente, os dados variantes no tempo são os que podem ser obtidos a partir do monitoramento do estudante na plataforma educacional de um AVA.

Existem diversos estudos na educação a distância que focam na predição de desempenho de alunos a distância em uma disciplina utilizando logs obtidos de AVAs. No trabalho de Zafra e Ventura (2009) são preditas notas de uma disciplina em um AVA onde são considerados o tempo dedicado, o número e os tipos de atividades realizadas pelos estudantes. No trabalho de Kotsiantis *et al* (2010) são utilizados combinações de classificadores a partir de Redes Bayesianas e algoritmos de redes neurais 1-NN e WINNOW usando uma metodologia de votação. É prevista a nota final de uma disciplina de um curso a distância, considerando as notas de quatro atividades. Em Saiz e Zorrila (2011) são comparados algoritmos de classificação para predizer o desempenho de um aluno em uma disciplina a distância. No trabalho de Lykourentzou *et al* (2008) é proposto um método de predição de evasão em cursos à distância utilizando redes neurais e dados extraídos de duas disciplinas de um AVA.

Todos os trabalhos mencionados focam na previsão de desempenho de apenas uma disciplina de um curso a distância em um AVA o que não possibilita fazer uma associação direta entre o insucesso em apenas uma disciplina e no curso como um todo. Ou seja, nem sempre um aluno que reprova determinada disciplina evadirá do curso de graduação à distância. Acreditamos que é importante identificar primeiramente quais são as disciplinas que mais influenciam na evasão do curso. Depois, utilizar essas disciplinas para predição de desempenho dos alunos ao longo das semanas. Acreditamos que com a predição de desempenho antecipada nessas disciplinas, será evitada a evasão do curso de graduação a distância. Assim, a presente abordagem propõe a utilização de Árvores de Decisão para predizer a evasão de um aluno em um curso de graduação a distância, identificando primeiramente quais disciplinas mais influenciam na evasão. Depois, poderão ser utilizadas essas disciplinas na predição de desempenho em um AVA.

2. Arquitetura de Identificação de Evasão proposta

A arquitetura de identificação de evasão de um curso a distância de graduação de maior duração proposta utiliza apenas dados variantes no tempo, pois pretende ser construída de forma automática, podendo assim, ser extensível para diferentes contextos. Na Figura 1 é vista a arquitetura da abordagem proposta.

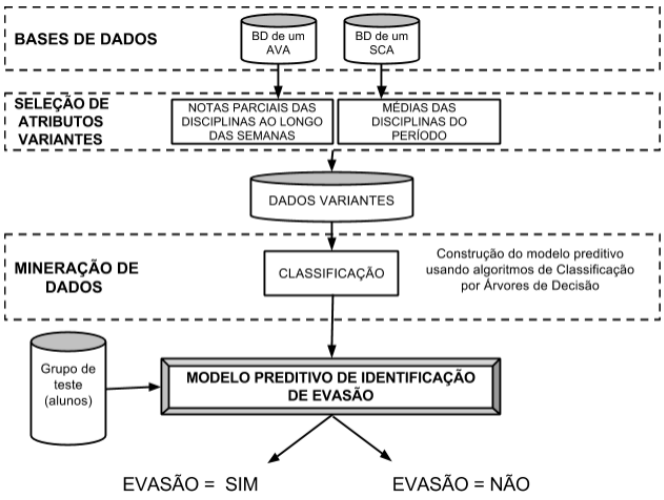

Figura 1. Arquitetura da Abordagem Temporal de Predição de Evasão Proposta

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 *Seleção de Atributos Variantes* é realizada a partir de pré-processamento e transformação dos dados composta por duas etapas: seleção das notas parciais das disciplinas obtidas da base de dados de um AVA e a seleção das médias das disciplinas da base de dados do SCA. No AVA são utilizadas as notas intermediárias das atividades das disciplinas do período ou semestre letivo ao longo das semanas para prever o desempenho final do aluno (aprovado ou reprovado), aqui denominados de "*Modelos na*". No SCA são utilizadas as médias finais das disciplinas que permitem a previsão de um aluno com risco de evasão ao final do período, aqui denominados de "*Modelos nb*". A etapa de *Mineração de Dados* com a aplicação dos algoritmos de Classificação para a construção dos modelos preditivos é mostrada nas próximas seções.

2.1. Abordagem Temporal a partir de Dados de um AVA (*Modelos na*)

Os *Modelos na* correspondem ao processamento de logs da disciplina de um AVA, que no presente estudo utiliza o Moodle. São propostas as seguintes tabelas do Moodle: *mdl_user_students*, *mdl_log* e *mdl_grades* onde são obtidas as notas parciais dos estudantes que são agrupadas de forma temporal por semanas. Cada atividade corresponde a uma tarefa realizada pelo aluno que possui uma nota. Na Tabela 1 é vista como as notas são organizadas para que seja possível a predição do desempenho do estudante.

Tabela 1. Modelo proposto para prever desempenho do estudante em um AVA - "*Modelo na*"

semana	1	2	3	m	média	Situação final da disciplina w
Notas do aluno 1	9	7	5	10	7,75	Aprovado
Notas do aluno 2	2	3	8	6	4,75	Reprovado
Notas do aluno 3	1	3	7	4	3,75	Reprovado
Notas do aluno n	7	5	9	7	7	Aprovado

Cada linha da Tabela 1 mostra as notas acumuladas nas semanas (1, 2, 3 e m) do aluno 1 ao aluno n, a média final do aluno e a situação final na disciplina. A partir da utilização de algoritmos de Classificação com as notas das atividades é possível prever antecipadamente se um aluno será aprovado ou reprovado na disciplina e a influência que cada semana tem no resultado final. No momento atual do trabalho foram

selecionadas bases de dados do AVA do Moodle para a realização dos experimentos e realizada extração de dados do AVA de apenas uma disciplina, conforme experimentos das seções 4.3 e 4.4.

2.2.Abordagem Temporal a partir de Dados de um SCA (*Modelos nb*)

Na Tabela 2 são mostradas as variáveis propostas utilizadas aos finais dos períodos ou semestres. 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. Todos os atributos são mostrados na Tabela 2.

Tabela 2. Variáveis utilizadas ao final do período obtidas do SCA

Variável	Valores
Situação da disciplina	(aprovado, reprovado por nota, reprovado por falta ou indefinido)
Média 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
Variável de Classe	Evadido ou graduado

A base de dados foi composta pelas médias das disciplinas e a situação final em cada uma delas, a média no período e, por fim, o atributo identificador da classe de aluno: graduado ou evadido, considerando o curso como um todo.

2.3.Integração entre os *Modelos na* e os *Modelos nb*

A Abordagem Temporal de identificação de estudante com risco de evasão aqui proposta faz a integração dos *Modelos na* e *Modelos nb*, pois a partir da descoberta de quais disciplinas mais influenciam na evasão a partir da aplicação dos *Modelos nb,* poderão ser priorizadas a previsão do desempenho dessas disciplinas nos *Modelos na*. A abordagem é aplicada em duas etapas:

1) Primeiramente são aplicados os *Modelos nb* para identificar quais são os atributos mais influentes na evasão ou graduação de um aluno considerando o curso como um todo.

2) Com a descoberta das variáveis mais importantes obtidas pelos *Modelos nb*, são obtidas as disciplinas que mais influenciam na evasão do curso como um todo. Dessa forma, os *Modelos na* serão aplicados para essas disciplinas a fim de prever o desempenho dos alunos.

Na Figura 2 pode ser observada a abordagem temporal proposta.

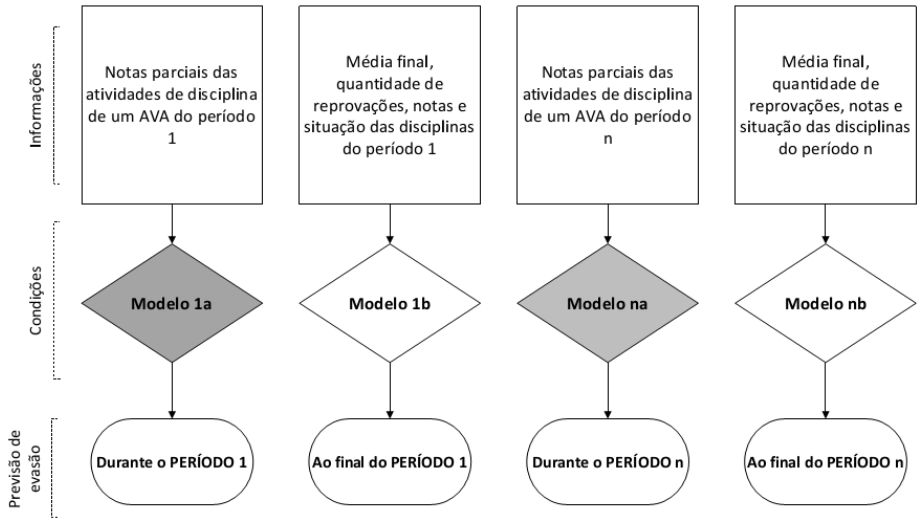

Figura 2. Arquitetura Temporal da Abordagem Preditiva de Evasão Proposta

Na Figura 2, as *informações* referem-se às entradas necessárias para a construção do modelo preditivo. A partir da entrada das informações em cada modelo, pode-se prever a aprovação ou reprovação em uma disciplina e a graduação ou evasão de um aluno no curso como um todo. Com o modelo preditivo construído, poderão ser antecipadamente identificados alunos com risco de evasão na fase de *condições*. O *Modelo 1a* refere-se ao modelo preditivo que prevê o desempenho de aluno em determinada disciplina do 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 determinada 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 a o desempenho de um aluno no 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 ao final do período para prever a evasão ao final do curso. Assim, medidas preventivas já poderão ser aplicadas para diminuir a probabilidade da ocorrência da evasão a partir da predição de desempenho das disciplinas identificadas nos *Modelos nb*. A característica principal da presente arquitetura temporal é investigar até que ponto outros períodos podem fornecer informações importantes sobre o risco de evasão de um aluno a partir da descoberta dos atributos que mais influenciam na evasão. Assim, o presente trabalho propõe, de forma inovadora, uma abordagem temporal para identificar um aluno com risco de evasão a partir da aplicação de vários modelos preditivos.

3. Metodologia e 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).

A escolha dos algoritmos pelas técnicas de Árvore de Decisão surgiu da necessidade de descobrir as disciplinas que mais influenciam na evasão do curso como um todo e também a influência dos demais atributos propostos no resultado final da

classe: evadido ou graduado. Árvores de Decisão são classificados como algoritmos de classificação por caixa branca, pois têm a capacidade de construir modelos que podem explicar as predições por regras do tipo: SE-ENTÃO, permitindo a explicação do funcionamento interno dos modelos que poderão ser utilizadas por atores da educação (gestores, professores, tutores, etc) nos mais diferentes propósitos.

4. Descrição das Bases de Dados Experimentos Realizados

Primeiramente, os experimentos foram realizados a partir da retirada do SCA da Unidade de Educação a 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 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 12 períodos (tempo suficiente para observar a evasão ou graduação de um aluno). O tempo normal de conclusão do referido curso 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 e 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 2) 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.

Depois foi realizado um experimento usando a disciplina identificada no *Modelo 1b* como a que mais influencia na evasão, *Introdução a Educação à Distância* (IED), do curso de Letras da UFPB Virtual. A base de dados foi dividida em duas classes distintas: os alunos aprovados e os reprovados.

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.

Em todos os experimentos realizados foram utilizados os seguintes algoritmos de Classificação por Árvore de Decisão: o SimpleCart (SC), o J48 (J48) e o ADTree (AT). As bases de dados foram divididas em 10 conjuntos utilizando o método da validação cruzada (10 fold cross-validation). Os algoritmos, aplicados a base de dados, foram executados 10 vezes, valor padrão de configuração do ambiente.

Nos experimentos realizados no SCA e no AVA foram aplicados filtros que identificaram quais atributos têm mais impacto na variável preditiva final: evadido ou graduado considerando o curso de graduação e aprovado ou reprovado considerando uma disciplina de um AVA. 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.

4.1 Experimentos utilizando todos os atributos na base de dados de um SCA (*Modelos nb*)

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. 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. 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 a média das acurácias obtidas em cada rodada dos classificadores mostrada na Tabela 3.

Tabela 3. 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 59 405	484 98 84 380	456 126 64 400

Na Tabela 3 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 SC: 447 alunos foram corretamente classificados como evadidos, 405 alunos foram corretamente classificados como graduados, 59 alunos foram incorretamente classificados como evadidos (de fato, se graduaram) e 135 alunos foram incorretamente classificados como graduados (de fato, evadiram o curso). 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 SC gerou uma árvore menor com apenas 6 nodos, conforme mostrado na Figura 3.

Figura 3 - Árvore de Decisão obtida com a aplicação do algoritmo SimpleCart

```
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
```

A Figura 3 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 corretamente classificados nesta regra e 39 alunos incorretamente classificados). Já para os alunos que tiraram média no primeiro período maior do que 5,36, outros atributos influenciam na graduação ou evasão do aluno.

Diante das árvores com muitos nodos resultantes dos algoritmos J48 e ADTree, surgiu a necessidade de utilizar filtros para escolha dos atributos mais relevantes para encontrar árvores de decisões com quantidade de nodos menores, o que possibilita a geração de regras mais simples, sendo assim, realizados os experimentos da seção 4.2.

4.2 Experimentos usando filtros para escolha dos atributos mais relevantes na base de dados de um SCA (*Modelos nb*)

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. A ferramenta calculou as médias das acurácias obtidas em cada rodada dos classificadores mostradas na Tabela 4.

Tabela 4. 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 63 401	459 123 79 385	429 153 51 413

4.3 Experimentos na base de dados de um AVA utilizando todos os atributos (*Modelos na*)

Foi identificada a disciplina 5 como a mais importante para predizer um aluno com risco de evasão, conforme experimentos mostrados nas seções 4.1 e 4.2. A disciplina 5 é a disciplina *Introdução da Educação a Distância (IED*). A base de dados dessa disciplina é composta por 38 alunos: 21 aprovados e 17 reprovados. Foram consideradas as 10 semanas iniciais da disciplina e a situação final: aprovado ou reprovado ao final da disciplina. Cada semana é considerada como a soma acumulativa das notas das atividades da disciplina na semana. A disciplina IED tem duração de 14 semanas, consistindo de atividades no ambiente Moodle que valem até 2000 pontos e a prova presencial que vale 1000 pontos. A média da disciplina é obtida a partir da soma dos pontos divididos por 3. Caso o resultado seja maior ou igual a 500 pontos, o aluno é considerado "APROVADO", caso seja menor do que 500 pontos, o aluno é considerado "REPROVADO". A ferramenta Weka calculou a média das acurácias obtidas em cada rodada dos classificadores (até determinada semana) onde foram calculadas as acurácias dos referidos algoritmos de Classificação das semana 1 a semana 10, conforme Tabela 5.

Tabela 5. Acurácias obtidas por semana da disciplina IED

Até a Semana	SC	J48	AT
1	60,52	71,05	65,78
2	65,78	68,42	68,42
3	65,78	71,05	78,94

4	68,42	76,31	71,05
5	71,05	81,57	76,31
6	71,06	81,57	73,68
7	89,47	84,21	78,94
8	89,47	84,21	78,94
9	89,47	81,57	78,94
10	89,47	81,57	78,94

É possível perceber que já na primeira semana se pode obter uma acurácia de 71,05% utilizando o algoritmo J48. Para facilitar a etapa de aplicação dos algoritmos de mineração de dados, todos os valores foram multiplicados por cem para evitar "vírgulas". A melhor acurácia foi obtida pelo algoritmo SC, na sétima semana, com 89,47%. Na base de dados, as notas da semana 2 e semana 7 possuem valor acumulado máximo de 150 pontos. Na Figura 4 é mostrada a árvore de decisão gerada pelo algoritmo SC considerando as primeiras dez semanas.

```
=== Classifier model (full training set) ===

CART Decision Tree

semana7 < 81.0
|  semana2 < 134.5: REPROVADO(15.0/0.0)
|  semana2 >= 134.5: APROVADO(2.0/1.0)
semana7 >= 81.0: APROVADO(19.0/1.0)
```

Figura 4. Árvore de Decisão gerada pelo algoritmo SC nas primeiras sete semanas aplicada à disciplina IED

Pela análise de Tabela 5, é possível perceber que as acurácias variam minimamente a partir da sétima semana (só há variação no algoritmo J48) o que sugere que as semanas 8,9 e 10 não possuem grande relevância no resultado da predição. Para confirmar esse resultado, foram utilizados filtros para escolha dos atributos mais relevantes, conforme mostrado na seção 4.4.

4.4 Experimentos na base de dados de um AVA usando filtros para escolha dos atributos mais relevantes aplicada à disciplina IED (*Modelos na*)

Todos os filtros selecionaram as semanas 1, 5 e 7, conforme análise das acurácias da Tabela 5, nesta ordem, como os mais relevantes. Quando aplicados os mesmos algoritmos, considerando apenas as semanas 1, 5 e 7 foram obtidas as seguintes acurácias, vistas na Tabela 6.

Tabela 6. Acurácias obtidas nas semanas determinadas pelos filtros da disciplina escolhida no experimento

Semana	SC	J48	AT
1	60,52	71,05	65,78
1 e 5	68,42	81,57	76,31
1,5 e 7	89,47	84,21	81,57

Apenas com os atributos da primeira, quinta e sétima semanas foram obtidas as mesmas acurácias que as obtidas por todas as semanas. Dessa forma, é interessante

utilizar filtros para escolher quais semanas mais influenciam na aprovação ou reprovação do aluno.

5. Conclusões e Trabalhos Futuros

A abordagem proposta têm muitas vantagens. A partir da aplicação dos *Modelos nb* é possível a identificação das disciplinas que mais estão relacionadas com à evasão. Dessa forma, pode-se dar maior ênfase nessas disciplinas a partir da realização de atividades de orientação e acompanhamento dos estudantes com o objetivo de prevenir a reprovação das mesmas, culminando com uma possível evasão do curso como um todo. A partir da aplicação dos *Modelos na* é possível que o professor possa identificar qual semana da sua disciplina é mais crítica no que se refere à evasão discente. Dessa forma, ele pode identificar os assuntos relacionados desta semana e focar em novas atividades de apoio ao estudante.

A partir dos experimentos realizados foi constatado que já ao final do primeiro período é possível prever o risco de um aluno evadir do curso de graduação a distância com acurácia maior do que 80%. Como trabalhos futuros serão realizados experimentos com os demais *Modelos na* e *Modelos nb* propostos. 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 para verificar se os resultados se repetem. A maior contribuição deste trabalho é propor uma arquitetura temporal de predição de evasão de um curso a distância de maior duração que identifica as disciplinas que mais influenciam na evasão, a partir da utilização de apenas dados variantes no tempo. O presente trabalho também mostrou que com apenas atributos variantes pode-se fazer a previsão de alunos com risco de evasão.

6. Referências

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.

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, 2010, p. 529-535.

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

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), p.147-165,.

SAIZ, D.G e ZORRILA,M. **Comparing classification methods for predicting distance students' performance**. JMLR: Workshop and Conference Proceedings 17. 2ns Workshop on Applications of Pattern Analysis (2011), p.26-32

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

ZAFRA, A e VENTURA, S. **Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming**. Educational Data Mining. Computer Science and Numerical Analysis Departament, University of Cordoba. (2009).