Previsão de Desempenho de Estudantes usando o Algoritmo de Classificação Associativa

Warley Leite Fernandes[1], Cristiano Grijó Pitangui[2], Alessandro Vivas Andrade[1], Luciana Pereira de Assis[1].

[1]Universidade Federal dos Vales do Jequitinhonha e Mucurí (UFVJM)

[2]Universidade Federal de São João Del Rei (UFSJ)

warleylfernandes@gmail.com, pitangui.cristiano@gmail.com, prof.alessandrovivas@gmail.com, lupassis@gmail.com

Abstract. *The present research focused on the extraction of data knowledge based on the AVA moodle databases of EAD. Aiming to identify students with potential for avoidance. Thus, we chose the CBA algorithm, because in this type of approach, it was not applied. The experimental results show that CBA is an excellent algorithm to generate classification rules and predict performance in educational bases, as it achieved better results than the traditional Classification algorithms, reaching an average accuracy of 83%. Additionally, results show that the forum, quiz and folder tools have a great influence on student performance.*

Resumo. *A pesquisa abordou a extração de conhecimento de dados com base nos bancos de dados do AVA moodle da EAD para identificar alunos com potencial para evasão. Utilizou-se o conceito de Séries Temporais, e o algoritmo CBA em conjunto com Predictive Apriori, que, entre as pesquisas realizadas, não havia sido empregada. Os resultados experimentais mostram que o CBA é um excelente algoritmo para gerar regras de classificação e prever desempenho em bases educacionais, pois atingiu melhores resultados que os algoritmos de Classificação tradicionais, alcançando uma acurácia média de 83%. Adicionalmente, resultados mostram que as ferramentas fórum, quiz e folder têm influência no desempenho dos estudantes.*

1. Introdução

A Educação a Distância (EAD) tem-se confirmado como importante ferramenta de capacitação a qualquer tempo e distância. Porém, a maioria das Instituições de Ensino tem encontrado dificuldades relacionadas ao grande número de evasões de seus cursos, que chegam, em média, a 40% (ABED 2016). Portanto, a evasão e a reprovação são, de forma geral, preocupações presentes na educação.

Na EAD esses problemas são potencializados devido às características específicas da modalidade, como: a falta de tempo por parte dos discentes, questões profissionais, falta de adaptação com o ambiente, problemas pessoais, insatisfação com o tutor, problemas técnicos, falta de suporte, dentre outros.

Avanços recentes em diversas áreas da tecnologia possibilitaram o surgimento das Tecnologias da Informação e Comunicação (TICs) que se tornaram essenciais à condução dos processos educacionais. Assim, grandes volumes de dados são gerados pela interação de usuários em Ambientes Virtuais de Aprendizagem (AVA). Existe o

DOI: 10.5753/cbie.wcbie.2017.734

desejo de que informações de grande valia aos professores e gestores acadêmicos possam ser obtidas das bases de dados educacionais, apesar de ser um processo de elevada complexidade. Neste sentido, faz-se necessário analisá-los de forma a descobrir conhecimento com objetivo de auxiliar na prevenção de problemas relacionados à evasão.

Segundo Silva *et. al.,* (2013), o processo de KDD ou Descoberta de Conhecimento em Bases de Dados é o ramo da computação que utiliza ferramentas e técnicas computacionais com a finalidade de sistematizar o processo de extração de conhecimento útil de grandes volumes de dados. O foco do KDD é, portanto, encontrar conhecimento significativo em Bases de Dados. Neste sentido, uma solução promissora para extração de informação é a Mineração de Dados, que atua iterativamente através do processo KDD.

A área de Mineração de Dados Educacionais (do inglês, *Education Data Mining* - EDM) foi criada da necessidade de se encontrar informações valiosas em Bases de Dados Educacionais. Neste sentido, um de seus desafios é auxiliar os professores no acompanhamento dos estudantes, devido à distância física que geralmente ocorre entre ambas as partes.

Diferentemente dos trabalhos pesquisados, este trabalho objetiva assinalar/compreender os motivos do baixo desempenho dos alunos em cursos EAD aplicando, com esta finalidade, o algoritmo de Classificação Associativa (CBA) (Agrawal *et. al.,* 1993). Existem diversas pesquisas relacionadas ao tema, como podem ser vistas em: (Romero *et. al.,* 2010) e (Rodrigues *et. al.,* 2014). Porém, a utilização do algoritmo CBA em EDM, com objetivo proposto, ainda não foi explorada.

Este trabalho optou pela utilização do CBA motivado pelos seguintes fatores:

- O CBA une duas técnicas muito utilizadas em EDM (Rodrigues *et. al.,* 2014), a saber: as Regras de Associação, e as Regras de Classificação.
- De forma geral, o CBA obtém resultados competitivos e às vezes superiores em relação aos algoritmos tradicionais de classificação quando aplicado a bases de dados não educacionais (Nofal *et. al*., 2010; Liu *et. al.,* 2001).
- A simplicidade das regras geradas pelo CBA possibilita uma maior clareza na compreensão do modelo extraído.
- O CBA foi pouco aplicado em EDM. Pesquisas realizadas apontaram apenas uma referência, (Badr *et. al.,* 2016), em que o mesmo foi utilizado com objetivo de prever a "taxa" de sucesso dos alunos em uma disciplina tendo como base suas notas em disciplinas anteriores.

As Bases de Dados utilizadas no presente trabalho advêm de um AVA *moodle* dos cursos Agente Comunitário de Saúde e Técnico em Hospedagem do Instituto Federal do Norte de Minas Ferais (IFNMG). Elas armazenam as interações dos alunos no AVA. O CBA foi avaliado em relação a outros algoritmos tradicionais de classificação, a saber: *BayesNet*, IBk, J48, *Random Forest*, JRip, *Multilayer Perceptron* e SVM, todos implementados no framework *Weka* (Eibe *et. al.,* 2016).

Resultados experimentais apontam que o CBA atinge melhores resultados em relação a algoritmos de classificação tradicionais. A exemplo, tem-se que o CBA obteve

em média 83% de acurácia na base de dados Agente Comunitário de Saúde, enquanto que o segundo melhor resultado foi atingido pelo *Multilayer Perceptron,* com acurácia média de 80%. Adicionalmente, resultados mostram que o uso das ferramentas *fórum, quiz, e folder* têm grande influência no desempenho dos estudantes, uma vez que estas ferramentas foram encontradas nas principais Regras de Associação geradas pelo CBA.

Este trabalho se organiza como segue. A seção 2 apresenta o referencial teórico necessário ao entendimento desta pesquisa. A seção 3 apresenta os materiais e métodos para realização deste trabalho. A seção 4 apresenta e discute os resultados experimentais realizados. Por fim, a seção 5 apresenta as conclusões e sugestões de trabalhos futuros.

2. Referencial Teórico

2.1 Regras de Associação

Segundo Larose (2005), as Regras de Associação, ou análise de afinidades, é o estudo dos atributos ou categorias de itens que estão relacionados, ou seja, àqueles que possuem alguma associação entre si. Esse método procura encontrar associações entre atributos, que configurem uma forma qualificada de relação entre eles.

Formalmente, uma Regra de Associação possui a seguinte estrutura: Se **A** então **B**, cuja representação é **A → B**, onde **A** é o antecedente e **B** o consequente da regra.

A qualidade de uma Regra de Associação é mensurada pelas medidas de *confiança* e *suporte*. De forma geral, dado um conjunto de transações (itens que podem estar relacionados), o objetivo da mineração de Regras de Associação é encontrar todas as regras que tenham valores de *confiança* e *suporte* iguais ou maiores que valores mínimos pré-estabelecidos.

A Tabela 1 apresenta as fórmulas para cálculo dos valores de *suporte e confiança* para uma Regra de Associação em sua forma geral, **A → B**.

Tabela 1. Fórmulas para cálculo de suporte e confiança

Fórmulas	Explicação
$\text{Suporte} = \frac{(A \cup B)}{T(A)}$	*Suporte* de uma regra **A → B** onde A e B são conjuntos de itens**.** O numerador se refere ao número de transações em que A e B ocorrem concomitantemente em um conjunto de dados e o denominador é o total de transações que o item A acontece.
$\text{Confiança} = \frac{(A \cup B)}{T}$	*Confiança* de uma regra **A → B** onde A e B são conjuntos de itens. O numerador se refere ao número de transações em que A e B ocorrem concomitantemente em um conjunto de dados e o denominador é o total de transações que acontece no conjunto de dados.

A Tabela 2, (Agrawal *et. al.,* 1993), ilustra um exemplo de transações de cestas de compras, onde cada linha representa uma transação (rotulada por um identificador único - TID), e um conjunto de itens comprados por um determinado cliente.

Tabela 2. Transações de cestos de compras

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Coca}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Coca}

A mineração de Regras de Associação deve encontrar regras consideradas fortes, onde o *suporte e a confiança* são maiores ou iguais aos valores pré-estabelecidos pelo usuário. Na tabela 2, por exemplo, pode-se "minerar" a Regra de Associação *Pão → Leite*, com valores de *suporte* e *confiança*, respectivamente, de 60% e 75%. Este valor de *confiança* aponta que em 60% do número total de transações, os itens Pão e Leite

foram comprados juntos. Por sua vez, o valor de *confiança* de 75% aponta que 75% dos clientes que adquiriram Pão também adquiriram Leite. Dessa forma, a regra, Pão → Leite, pode ser considerada forte, uma vez que os resultados do *suporte* e da *confiança* são maiores ou iguais aos valores preestabelecidos, e, por esse motivo, pode ser considerada como informação valiosa e deve ser analisada com maior atenção.

2.2 Os Algoritmos *Apriori* e *Predictive Apriori*

O primeiro algoritmo proposto para a mineração de Regras de Associação foi o *Apriori* (Agrawal *et. al.,* 1993), que ainda hoje é um dos mais usados e considerado um clássico quando se lida com o problema de extração de Regras de Associação. Para gerar as Regras de Associação, o *Apriori* possui como parâmetros de entrada os valores de *confiança* e *suporte*. Assim, o *Apriori* busca por regras que possuam valores maiores ou iguais aos parâmetros de *confiança e suporte* fornecidos pelo usuário.

O *Predictive Apriori* (Scheffer 2001), que se baseia no algoritmo *Apriori*, combina as métricas de *confiança e suporte* em uma medida única, chamada de *acurácia preditiva*. O algoritmo busca por Regras de Associação ordenando-as segundo esta métrica. De forma geral, a o mecanismo de busca do *Predictive Apriori* aumenta gradativamente os valores de *confiança e suporte* para encontrar suas Regras de Associação. Portanto, o *Predictive Apriori* trabalha no sentido de buscar regras mais equilibradas em relação aos valores de *confiança* e *suporte* (Garcia *et. al.,* 2006) quando comparado ao algoritmo *Apriori.*

2.3 Regras de Associação em EDM

A mineração de Regras de Associação em Bases de Dados Educacionais pode ser aplicada com diversas finalidades, tais como: Análise do Perfil de Aprendizagem, Análise de Perfil de Comportamento, Análise de Desempenho, dentre outras. Alguns dos trabalhos mais relevantes relacionados a esta pesquisa são brevemente descritos a seguir.

Nunes *et al.* (2015) têm como objetivo explorar a aplicabilidade de técnicas de EDM em Mundos Virtuais. Para tanto, um estudo de caso foi desenvolvido em um laboratório de química, o qual contém diferentes tipos de atividades educacionais, como a apresentação de *slides*, vídeos e atividades de cunho prático. Foram analisados possíveis padrões de comportamento na base de dados por meio da mineração de Regras de Associação com o uso do algoritmo *Apriori*, que possibilitou a identificação de eventuais mudanças no planejamento pedagógico das atividades.

Wilves *et al.* (2015) têm como objetivo validar em um novo contexto Regras de Associação já mineradas, a fim de evidenciar o desânimo do aluno quando realiza atividades individuais e em grupo. A partir da adaptação de algumas Regras de Associação, foi possível obter um modelo genérico para a descoberta do padrão de comportamento do aluno desanimado, que pode ser usado como subsídio ao professor.

Santos *et al.* (2014) aplicam Regras de Associação para identificar em quais cursos os alunos tendem a ser retidos e quais cursos os alunos tendem a ser aprovados. Foi possível verificar a correlação entre cursos e retenção de alunos. A seleção e avaliação do conjunto de regras foi feita por especialistas (Chefe de cursos, Coordenador e Professores). Assim, foi possível propor algumas sugestões sobre a organização do fluxograma de alguns programas a fim de reduzir as taxas de retenção.

Penedo *et al.* (2012) estudam uma amostra de dados reais referentes ao *log* de acessos do AVA do Consórcio CEDERJ (Universidades Públicas a Distância) para identificar o padrão dos usuários que melhor se adaptam ao sistema de EAD disponibilizado. As regras descobertas por meio do *Apriori* apontam para uma tendência maior de utilização das ferramentas disponibilizadas pela plataforma que dizem respeito às disciplinas (materiais e exercícios complementares), sendo as ferramentas relacionadas a aplicativos (wiki, blog, fórum) pouco utilizadas. Outra descoberta apontam os horários de maior utilização da plataforma (tarde e noite).

2.4 O Algoritmo de Classificação Associativa (CBA)

O algoritmo *Classification Based on Association* (CBA) (Liu *et. al.*, 1998) integra duas técnicas de extração de conhecimento (Associação e Classificação) para construir classificadores mais precisos. Seu funcionamento consiste em duas etapas principais, a saber:

I. Realiza-se a extração de um subconjunto especial de Regras de Associação, baseando-se no algoritmo *Apriori*, cujos lados direitos das regras são restritos a classificação do atributo classe. Este subconjunto de regras é chamada de *class association rules* (CARs).

II. Constrói-se um classificador baseando-se nas CARs geradas. Esse classificador é uma lista de regras ordenadas de acordo com o valor de *confiança* fornecido pelo usuário.

No que diz respeito a Bases de Dados Educacionais, o algoritmo CBA foi pouco aplicado. Em pesquisas realizadas, encontrou-se apenas um trabalho, Badr *et. al.,* (2016), que aplica o CBA para prever o desempenho de estudantes na disciplina de Programação tendo como base suas notas em disciplinas já cursadas.

3. Materiais e Métodos

Atualmente, grande parte das Instituições Educacionais utiliza os AVAs como apoio ao ensino e aprendizagem dos alunos, tanto nos cursos presenciais, semipresenciais e principalmente à distância. O AVA no processo de ensino e aprendizagem proporciona a interação dos alunos e professores através inúmeros recursos disponibilizados, como: materiais (documentos e vídeos), *chats*, fóruns e outras ferramentas. A utilização destes recursos produz uma enorme quantidade de dados de difícil interpretação sem o uso de técnicas de EDM.

Esta pesquisa trata do problema de evasão dos cursos de Agente Comunitário de Saúde e Técnico em Hospedagem da EAD do Instituto Federal do Norte de Minas Gerais. Para isto, são utilizados atributos que representam o comportamento dos estudantes no AVA *moodle*. A Tabela 3 exibe o conjunto de atributos explorados.

3.1 Bases de Dados Utilizadas

Realizou-se uma análise detalhada das bases de dados do AVA do IFNMG dos cursos de Agente Comunitário de Saúde e Técnico em Hospedagem, com a finalidade de identificar os dados que têm relevância na caracterização do perfil dos estudantes. Dessa forma, os dados selecionados foram submetidos ao CBA com a finalidade de identificar os alunos que possuem características e comportamentos que podem levar à evasão ou à reprovação.

Com o objetivo de melhor caracterizar o problema abordado, as notas dos alunos foram discretizadas. Para isso, as bases de dados foram divididas em duas classes (Aprovado e Reprovado). Os alunos foram divididos de acordo com suas situações finais no curso, onde a classe Aprovado é representada pelos alunos com nota maior ou igual a 60, e a classe Reprovado é representada pelos alunos que obtiveram nota inferior a 60. O restante dos atributos foi discretizado pelo método *equal-width* (Dougherty *et. al.*, 1995) com três intervalos e rótulos (baixo, médio e alto) (Avlijas *et. al*., 2016). Optou-se pela discretização de atributos, já que alguns algoritmos utilizados não trabalham com dados contínuos. Durante as etapas de pré-processamento e transformação dos dados, foram separados 27 atributos dos estudantes do AVA *moodle* juntamente com o atributo classe. A tabela 4 apresenta um resumo dos conjuntos de dados obtidos após a etapa de pré-processamento.

Tabela 3. Conjunto de Atributos Explorados

Grupo	Atributos	Descrição
Fórum	nr_forum	Número de acesso à página principal do Fórum.
	add_post_forum	Adição de post na discussão do Fórum.
	search_forum	Fez uma pesquisa no Fórum.
	forum_discussion	Acessou a discussão no Fórum.
	forum_view	Acessou o Fórum.
	up_forum	Atualizou o acesso ao Fórum.
	del_forum	Quantidade de postagens apagadas de uma discussão.
	subscribe_forum	Inscreveu-se no Fórum.
	report_forum	Acesso ao relatório do usuário no Fórum.
Questionário (quiz)	nr_quiz	Número de acesso ao questionário.
Chat	nr_chat	Número de acesso à página principal do *Chat*.
	chat_view	Acesso à sala do *Chat.*
	report_chat	Acesso aos relatórios.
	talk_chat	Mensagens enviadas no *Chat.*
Folder	nr_folder	Número de acesso à pasta de conteúdo.
	folder_view	Acessou o conteúdo da pasta.
Usuários (user)	nr_acesso_user	Acesso à página principal dos usuários.
	user_view	Acesso a determinado usuário.
	user_view_all	Acesso à página com todos os usuários.
Ambiente Virtual	nr_login_AVA	Acesso ao AVA.
Curso	course_view	Acesso ao curso.
Recurso	resource_view	Visualização de apostilas ou vídeos disponibilizados.
Informações	url_view	Acesso à página de informações do curso (ementa, bibliografia...).
Blog	blog_view	Acesso ao *blog*.
Tarefas (formulários)	assign_submit	Envio de atividades.
	assign_view	Acesso às atividades do curso.
	assign_submit_form	Acesso ao formulário de submissão atribuído.

Tabela 4. Relação dos conjuntos de dados obtidos e suas respectivas classes

Base de Dados	Nº Estudantes	Nº de estudantes na classe Aprovado	Nº de estudantes na classe Reprovado
Agente Comunitário de Saúde	112 (100%)	47 (41,96%)	65 (58,04%)
Técnico em Hospedagem	111 (100%)	54 (48,65%)	57 (50,45%)

3.2 Ferramentas Utilizadas

O software *Weka* (Eibe *et. al.*, 2016) foi utilizado nas etapas experimentais deste trabalho. Optou-se pelos seguintes algoritmos de classificação: IBk, J48, *Random Forest* (RF), JRip, *Multilayer Perceptron* (MP). Optou-se também pelos algoritmos de Regras de Associação *Apriori* e *Predictive Apriori*, e pelo algoritmo de CBA que é a base de desenvolvimento desta pesquisa.

3.3 Metodologia Experimental

Nos cursos do IFNMG usados neste estudo, as disciplinas são ofertadas no formato semestral e divididas em 4 módulos. No fim de cada módulo, é realizada uma avaliação, e no fim do semestre ocorre um encontro presencial, onde os estudantes são avaliados. Devido a esta dinâmica, os experimentos realizados neste trabalho fizeram o

uso do conceito de séries temporais. Dessa forma, o período de oferta da disciplina foi dividido em 4 sub-períodos com duração máxima de 45 dias. Como a oferta da disciplina foi dividida em 4 períodos de tempo, foram gerados 4 bases de dados para cada uma das bases de dados iniciais. Cada conjunto possui as interações dos estudantes com o AVA até o corte temporal.

A divisão do conjunto de dados em períodos temporais se justifica, uma vez que, dessa maneira, os professores e tutores (responsáveis pela turma) poderão realizar um acompanhamento progressivo dos estudantes. Para avaliação do desempenho dos algoritmos de classificação utilizou-se a métrica Acurácia (Ac.) e *F-Measure (F),* esta última para avaliar o desempenho de predição do classificador em cada classe individualmente.

Adotou-se o método de *10-fold cross-validation* (Mitchel, 1997) para treinamento e parametrização dos algoritmos de classificação. O algoritmo CBA foi configurado de duas formas: utilizando o *Apriori* e o *Predictive Apriori*. Os parâmetros usados pelo algoritmo *Apriori* para geração das Regras de Associação foram *suporte* = 25% e confiança = 75%. Os parâmetros dos algoritmos de classificação foram J48: confidencefactor = 0,3 (valor de confiança para poda), minNumobj = 2 (número mínimo de instâncias por folha (nó)); Random Forest: bagSizePercent = 100; Jrip: folds=10 (determina a quantidade de dados para poda); IBK: KNN=1 (determina o número de vizinhos usados); SVM: SVM – Type = nu – svc (tipo de classificador), KernelType = linear:u´*v (kernel base); RandomForest: ValidationSetSize = 25 (tamanho da porcentagem do conjunto de validação), learningreate = 0,3 (taxa de aprendizagem); BayesNet: estimator=SimpleEstimator (Algoritmo de estimação para encontrar a tabela de probabilidade condicional a rede bayesiana) search algorithm = k2 (método usado para pesquisa das estruturas das redes). Em todos os algoritmos foram configurados o parâmetro batchsize (define o número de instâncias usadas para predição), nesse caso o valor configurado variou entre 100 e 120 de acordo com cada base minerada.

4. Experimentos

A Tabela 5 apresenta os resultados dos algoritmos aplicados às séries temporais utilizando o conjunto de dados dos cursos Técnico em Hospedagem e Agente Comunitário de Saúde.

Resultados apontam que o CBA + *Predictive Apriori* obtém melhores resultados em comparação ao CBA + *Apriori*, onde alcançou uma acurácia média de 83% na base de dados Agente Comunitário de Saúde e 78,5% na base Técnico em Hospedagem. Observou-se também que os valores da F-Measure da classe Aprovado são inferiores aos valores da classe Reprovado, usando o algoritmo CBA com o *Apriori*. Ao aplicar o algoritmo CBA com o *Predictive Apriori,* percebeu-se um maior equilíbrio nos valores da F-Measure das duas classes, e consequentemente, um aumento de acertos de classificação das instâncias da classe Aprovado, melhorando os valores da F-Measure. Dessa forma, o CBA + *Predictive Apriori*, atingiu o maior valor de acurácia dentre todos os algoritmos avaliados, e também um maior equilíbrio entre as classes.

Podem-se observar indícios de queda na Acurácia dos classificadores gerados pelo CBA no decorrer dos períodos da série temporal. Apesar do aumento do volume de dados, que ocorre naturalmente no decorrer do curso, há uma redução do uso de algumas ferramentas na plataforma por diversos estudantes. Consequentemente, o

acréscimo dos dados se dá apenas em alguns atributos, e outros caem em desuso. Dessa forma, há uma queda de alunos classificados corretamente pelo algoritmo CBA, e consequentemente uma queda nos valores de acurácia, ou seja, menos atributos, menor a acurácia. Assim, podemos inferir que isso ocorre por causa do desuso de alguns atributos. Pois, de acordo com a tabela 6, alguns atributos, como por exemplo: *add_post_forum* e *search_forum*, que têm boa influência no desempenho dos estudantes de acordo com as suas regras no corte temporal 1, ficaram em desuso nas outras séries.

Tabela 5 – Resultados da Classificação

	Agente Comunitário de Saúde									Técnico em Hospedagem								
Algoritmo	**1**		**2**		**3**		**4**		**Média**	**1**		**2**		**3**		**4**		**Média**
	Ac.	*F*	*Ac.*	*F*	*Ac.*	*F*	*Ac.*	*F*	*Ac.*	*Ac.*	*F*	*Ac.*	*F*	*Ac.*	*F*	*Ac.*	*F*	*Ac.*
CBA com *Predictive Apriori*	**85,0%**	0,858 0,841 0,851	84,3%	0,866 0,812 0,846	82,9%	0,848 0,804 0,830	80,2%	0,841 0,738 0,798	**83%**	84,9%	0,860 0,837 0,848	81,3%	0,827 0,796 0,810	75,5%	0,761 0,748 0,754	72,7%	0,746 0,706 0,725	78,5%
CBA com *Apriori*	77,5%	0,834 0,648 0,757	82,0%	0,831 0,808 0,821	74,8%	0,816 0,600 0,726	74,8%	0,816 0,600 0,726	77,3%	78,2%	0,745 0,810 0,778	74,1%	0,719 0,760 0,739	71,8%	0,667 0,756 0,712	71,8%	0,667 0,756 0,712	74%
IBk	68,5%	0,752 0,568 0,676	73,0%	0,786 0,634 0,723	69,4%	0,767 0,553 0,678	70,3%	0,769 0,582 0,692	70,2%	68,2%	0,632 0,720 0,677	70,9%	0,680 0,733 0,707	70,9%	0,680 0,733 0,707	67,3%	0,673 0,673 0,673	69,3%
J48	74,8%	0,803 0,650 0,739	76,6%	0,819 0,667 0,756	69,4%	0,773 0,528 0,672	72,1%	0,783 0,608 0,710	73,2%	71,8%	0,710 0,726 0,718	64,5%	0,562 0,702 0,633	70,9%	0,610 0,768 0,690	70,9%	0,610 0,768 0,690	69,5%
JRip	62,2%	0,720 0,417 0,594	62,2%	0,727 0,382 0,584	58,6%	0,712 0,258 0,524	58,6%	0,705 0,303 0,538	60,4%	76,4%	0,717 0,797 0,758	75,5%	0,733 0,773 0,753	59,1%	0,494 0,656 0,577	65,5%	0,486 0,740 0,615	69,1%
MP	77,5%	0,806 0,731 0,775	81,1%	0,832 0,784 0,812	80,2%	0,828 0,766 0,802	81,1%	0,835 0,779 0,812	80%	73,6%	0,695 0,768 0,732	78,2%	0,778 0,786 0,782	70,0%	0,708 0,692 0,700	70,0%	0,708 0,692 0,700	73%
RF	70,3%	0,769 0,582 0,692	64,0%	0,744 0,394 0,599	61,3%	0,723 0,358 0,572	69,4%	0,770 0,541 0,675	66,3%	72,7%	0,694 0,754 0,725	73,6%	0,701 0,764 0,733	70,0%	0,673 0,723 0,698	70,0%	0,673 0,723 0,698	71,6%
SVM	77%	0,809 0,711 0,767	82,3%	0,863 0,752 0,822	80,5%	0,841 0,750 0,802	76,1%	0,787 0,727 0,762	78,9%	77%	0,595 0,750 0,674	78,2%	0,782 0,782 0,782	71,8%	0,705 0,730 0,718	71,8%	0,705 0,730 0,718	74,7%
BN	80,2%	0,843 0,732 0,797	81,1%	0,842 0,764 0,810	79,3%	0,830 0,736 0,791	79,3%	0,830 0,736 0,791	80%	73,6%	0,701 0,764 0,733	75,5%	0,710 0,787 0,749	72,7%	0,694 0,754 0,725	72,7%	0,694 0,754 0,725	73,7%

4.1 Discussão

A tabela 6 apresenta as Regras de Associação da base Agente Comunitário de Saúde gerado pelo CBA + *Predictive Apriori.* Percebe-se a simplicidade e a importância das regras geradas. Para ilustrar este fato, a primeira regra do corte temporal 1 na Tabela 6, é interpretada da seguinte maneira: se os alunos tiverem baixa interação nos atributos *nr_forum* e *nr_folder* serão reprovados. Ou seja, aqueles alunos que participam pouco do fórum e pouco acessam o material que fica disponível na pasta de conteúdo, tendem a ser reprovados. É importante notar o alto valor de Acurácia desta regra (95%).

As regras expostas na tabela 6 apresentam alta capacidade preditiva e, analisando-as, pode-se notar uma grande influência das ferramentas *fórum, quiz e folder* (com uso de seus atributos *nr_forum*, *nr_folder* e *nr_quiz*) no desempenho dos estudantes. Nota-se a presença desses atributos em cada corte temporal e com grande influência na Aprovação dos estudantes. Por exemplo: a regra, nr_quiz = alto → situação = Aprovado (A), onde o número de acesso ao questionário (atributo = nr_quiz) considerado como alto, define a situação do aluno como Aprovado. Tal regra apresenta-se em todas as séries temporais, e com Acurácia Preditiva de 91%, 91%, 90% e 86% respectivamente. Dessa forma evidencia a importância dessas ferramentas no resultado final do estudante.

Observou-se que no decorrer da série temporal há uma redução de uso de algumas ferramentas, e consequentemente, uma queda nos resultados dos

classificadores observados na Tabela 5. Nota-se, por exemplo, que os atributos *add_post_forum* e *search_forum*, que têm influência no desempenho dos estudantes na série temporal 1, simplesmente caíram em desuso nas outras séries. Outros atributos tiveram interações em períodos intercalados da série, como os atributos *user_view* e *forum_discussion*. A análise desta diversificação dos usos das ferramentas em diferentes momentos na disciplina pode revelar pontos interessantes que precisam ser investigados mais profundamente.

Tabela 6 – Principais regras de Associação encontradas pelo algoritmo CBA

Séries	Principais Regras de Associação Encontradas	Acp.
1	nr_forum=Baixa nr_folder=Baixa ➔ situação = REPROVADO (A)	95%
	add_post_forum=Alta ➔ situação = APROVADO (A)	91%
	nr_quiz=Alta ➔ situação = APROVADO (A)	91%
	nr_folder=Alta ➔ situação = APROVADO (A)	86%
	resource_view=Baixa nr_folder=Baixa nr_chat=Baixa Add_post_forum=Baixa user_view=Baixa ➔ situação = REPROVADO (A)	86%
	nr_chat=Alta ➔ situação = APROVADO (A)	84%
	forum_discussion=Baixa add_post_forum=Baixa folder_view=Baixa search_forum=Baixa ➔ situação = REPROVADO (A)	84%
2	nr_forum=Alta ➔ situação = APROVADO (A)	98%
	forum_view=Alta ➔ situação = APROVADO (A)	93%
	nr_quiz=Alta ➔ situação = APROVADO (A)	91%
	nr_forum=Baixa forum_view=Baixa ➔ situação = REPROVADO (A)	91%
	nr_forum=Baixa ➔ situação = REPROVADO (A)	89%
	resource_view=Média assign_view=Média ➔ situação = APROVADO (A)	88%
	nr_forum=Baixa course_view=Baixa ➔ situação = REPROVADO (A)	88%
3	nr_forum=Baixa course_view=Baixa nr_folder=Baixa ➔ situação = REPROVADO (A)	96%
	forum_discussion=Alta ➔ situação = APROVADO (A)	93%
	nr_quiz=Alta ➔ situação = APROVADO (A)	90%
	nr_forum=Baixa nr_login_AVA=Baixa ➔ situação = REPROVADO (A)	88%
	nr_forum=Baixa resource_view=Baixa nr_folder=Baixa nr_chat=Baixa ➔ situação = REPROVADO (A)	86%
	nr_forum=Alta ➔ situação = APROVADO (A)	86%
4	nr_forum=Baixa course_view=Baixa nr_folder=Baixa ➔ situação = REPROVADO (A)	96%
	nr_forum=Baixa nr_folder=Baixa ➔ situação = REPROVADO (A)	93%
	nr_login_AVA= Baixa courseview=Baixa nr_folder=Baixa ➔ situação = REPROVADO (A)	91%
	nr_forum=Baixa nr_login_AVA=Baixa ➔ situação = REPROVADO (A)	90%
	nr_forum=Alta ➔ situação = APROVADO (A)	86%
	nr_quiz= Alta ➔ situação = APROVADO (A)	86%
	course_view=Baixa nr_folder=Baixa nr_chat=Baixa user_view=Baixa ➔ situação = REPROVADO (A)	83%

5. Conclusão

Considerando demandas cada vez maiores da EAD, esta pesquisa aplicou técnicas de EDM em bases de dados adivindas de um AVA, com o objetivo de identificar alunos com potencial para evasão, e, com isso, ajudar a gestão acadêmica e pedagógica da instituição de ensino a tomar as decisões necessárias para evitar a desistência dos estudantes. Para isto, aplicou-se o CBA, um algoritmo poderoso, mas muito pouco utilizado em EDM.

Resultados experimentais apontaram que ferramentas tais como *fórum, quiz e folder* possuem grande impacto no desempenho dos estudantes. Adicionalmente, apontou-se que o algoritmo CBA, utilizado em Bases de Dados Educacionais, obtém resultados preditivos comparáveis aos principais algoritmos de classificação. Com os cortes temporais os gestores da turma podem ter um resultado progressivo dos estudantes e três *feedback* antes da avaliação final (que acontece no final do curso em um encontro presencial). Mostrou-se também a simplicidade de interpretação das regras de associação geradas pelo CBA, que pode auxiliar os profissionais de Informática na Educação a descobrir padrões e tendências implícitas nos conjuntos de dados.

Como trabalhos futuros, pretende-se realizar novos experimentos em mais bases dados educacionais, usando técnicas de balanceamento e seleção de atributos com o algoritmo CBA, para possível melhora dos resultados.

6. Referências

ABED. (2016). Anuário Brasileiro Estatístico de Educação Aberta e a Distância.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data, v. 22, n. May, p. 207–216.

AVLIJAS, G.; HELETA, M.; AVLIJAS, R. (2016). A guide for association rule mining in moodle course management system. p. 56–61.

BADR, G. et., al. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. v. 82, n. March, p. 80–89.

DOUGHERTY, J., KOHAVI, & SAHAMI, M. (1995). Supervised and unsupervised discretization of continuous features.

EIBE FRANK, MARK A. HALL, and IAN H. WITTEN. (2016). The WEKA Workbench. Morgan Kaufmann, Fourth Edition.

GARCÍA, E., Romero, C., Ventura, S., & de Castro, C. (2006). Using rules discovery for the continuous improvement of e-learning courses.

LAROSE, D. T. (2005). Discovering knowledge in data. [s.l: s.n.]. v. 53

LIU, B.; HSU, W.; MA, Y. (1998). Integrating Classification and Association Rule Mining. Knowledge Discovery and Data Mining, p. 80–86.

LIU, B.; MA, Y.; WONG, C. (2001). Classification using association rules:Weaknesses and enhancements.DataMining for scientific and engineering applications,p.591-605.

MITCHELL, T.M. (1997). Machine Learning. McGraw-Hill.

NOFAL, M.; BANI-AHMAD, S. (2010). Classification Based on Association-Rule Mining Techniquese a General Survey and Empirical Comparative Evaluation.

NUNES, F. B.; VOSS, G. B.; CAZELLA, S. C. (2015). Mineração de dados educacionais e Mundos Virtuais : um estudo exploratório no OpenSim.

PENEDO, J. R.; CAPRA, E. P. (2012). Mineração de Dados na Descoberta do Padrão de Usuários de um Sistema de Educação à Distância. n. Sbsi, p. 396–407.

RODRIGUES, R. L. et., al. (2014). A literatura brasileira sobre mineração de dados educacionais. Cbie. n. 3, p. 621–630.

ROMERO, C.; VENTURA, S. (2010). Educational data mining: A review of the state of the art. IEEE, v. 40, n. 6, p. 601–618.

SANTOS, M. S. et., al. (2014). Mining Retention Rules from Student Transcripts : A Case Study of the programs at a Federal University. n. Cbie, p. 762–771.

SILVA, C. V. A. et., al. (2013). Mining Retention Rules from Student Transcripts: A Case Study of the Information Systems programme at a Federal University. Cbie.

SCHEFFER, TOBIAS. (2001). "Finding association rules that trade support optimally against confidence".(PKDD-01).

WILVES, F. D. S. L. K.; CAZELLA, S. C. (2015). Analisando o desânimo de alunos em ambientes virtuais através da mineração de dados educacionais. p. 65–70.