Uso de Séries Temporais e Seleção de Atributos em Mineração de Dados Educacionais para Previsão de Desempenho Acadêmico

Rodrigo Magalhães Mota dos Santos[1], Cristiano Grijó Pitangui[2], Alessandro Vivas Andrade[1], Luciana Pereira de Assis[1]

[1] Universidade Federal dos Vales do Jequitinhonha e Mucuri

[2] Universidade Federal de São João del Rei

{rodrigo.mota, lpassis}@ufvjm.edu.br, alessandro.vivas@gmail.com, Pitangui.cristiano@ufsj.edu.br

Abstract. *The academic performance prediction can be very useful for Educational Institutions in order to help them to take pedagogical decisions that can help students. In this work, we present experiments using Moodle data, Time Series and the Feature Selection Wrapper approach, since, to best of our knowledge, this method to reduce the number of features have not been used in this "kind" of data. Results showed an improvement in the performance of classifiers, some obtaining the mark of 84.7% in accuracy results.*

Resumo. *A previsão de desempenho acadêmico tem grande utilidade para Instituições de Ensino no sentido de auxiliá-las a tomar, de forma antecipada, decisões pedagógicas que possam auxiliar os estudantes. Neste trabalho foram realizados experimentos em uma base de dados do Ambiente Virtual de Aprendizagem Moodle, utilizando o conceito de Séries Temporais e a técnica Cápsula de Seleção de Atributos que, dentre os trabalhos pesquisados, não havia sido ainda empregada. Resultados experimentais indicam uma melhora no desempenho dos classificadores com o uso de Seleção de Atributos, alguns alcançando a marca de 84,7% de acurácia.*

1. Introdução

A tecnologia, presente cada vez mais no ambiente educacional, tem contribuído para o aumento da oferta de cursos à distância. Grande parte dos cursos ofertados nesta modalidade utilizam os Ambientes Virtuais de Aprendizagem (AVA). Estes ambientes ganham espaço no cotidiano dos educadores devido ao fácil manuseio e a grande diversidade de ferramentas disponibilizadas. Tais ferramentas permitem, de forma geral, a administração de cursos totalmente à distância com oferta de múltiplas mídias e recursos (fóruns de discussão, *chats*, dentre outros) para interações entre professores e alunos.

Com tamanha gama de recursos, os AVA's se tornaram detentores de inúmeras informações, registros de acessos, e interações. Este enorme volume de dados armazenado está sendo estudado por profissionais da área de Informática na Educação (IE).

DOI: 10.5753/cbie.sbie.2016.1146

Com a utilização do processo de Descoberta de Conhecimento em Base de Dados (do inglês, *Knowledge Discovery in Databases* - KDD), profissionais da IE podem descobrir padrões e tendências implícitas em grandes conjuntos de dados. O processo de KDD preocupa-se com o desenvolvimento de métodos e técnicas para dar sentido aos dados. O objetivo do KDD é identificar relações entre os dados analisados para gerar novos conhecimentos e descobertas científicas. O núcleo do processo é a aplicação de métodos específicos de Mineração de Dados para a descoberta e extração de padrões compreensíveis, válidos, novos, e potencialmente úteis, a partir de grandes conjuntos de dados [Fayaad *et al*., 1996].

A Mineração de Dados (do inglês *Data Mining* - *DM*) é a área específica do KDD que trata das técnicas e algoritmos utilizados na detecção dos padrões de dados. Segundo Hand *et al.* (2011), DM é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados.

Com a expansão dos cursos mediados por meios computacionais, muitos pesquisadores na área da IE têm utilizado DM para explorar dados originados em ambientes educacionais, com o objetivo de encontrar padrões e contribuir com os processos de aprendizagem e, consequente, melhoria das ações pedagógicas e de materiais didáticos. Dentro deste contexto, surgiu a área de pesquisa conhecida como Mineração de Dados Educacionais (do inglês, *Education Data Mining* - EDM). Segundo Baker *et al.* (2011), EDM é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem.

O presente trabalho apresenta um estudo de algumas técnicas de EDM aplicadas em uma base de dados do AVA *Moodle*. O objetivo principal é a geração de estimativas de desempenho acadêmico, pois através destas estimativas pode-se analisar se um estudante tem tendência à aprovação ou reprovação em um curso. Neste trabalho é utilizado o conceito de Séries Temporais [Hall *et al*., 2011], que permite o acompanhamento estudantil durante toda a realização do curso. Os modelos de classificação propostos neste trabalho utilizam a inserção das notas de avaliações, à medida que vão sendo realizadas, e ainda utiliza a técnica de Seleção de Atributos em Cápsula [Kohavi *et al*., 1996]. Optou-se por esta abordagem, porque, dentre os trabalhos pesquisados, nenhum fez uso deste método em conjuntos de dados retirados de cursos de graduação à distância ofertado em AVA's, e também pelo fato deste método alcançar excelentes resultados em outros "tipos" de bases de dados [Han *et al.*, 2012; Mitchell, 1997].

Este trabalho se organiza em 5 seções. A seção 2 apresenta os trabalhos relacionados a esta pesquisa. A seção 3 apresenta os materiais utilizados e os métodos aplicados na realização deste trabalho. A seção 4 apresenta os experimentos realizados, seus resultados e suas respectivas análises. Finalmente, a seção 5 apresenta as conclusões deste trabalho e sugere algumas perspectivas de continuidade de pesquisa.

2. Trabalhos Relacionados

Existem diversos trabalhos que utilizam a tarefa de classificação para prever o desempenho acadêmico de estudantes, dentre eles, Ferreira (2015) tem como foco a

identificação de fatores relacionados à conclusão do Ensino Fundamental utilizando técnicas de mineração de dados aplicadas aos Micro-dados do Censo Escolar. Foi aplicado o algoritmo de classificação *J48* e o filtro *CfsSubsetEval,* ambos da ferramenta *Weka* [Bouckaert *et al*., 2015]. Os resultados apresentaram uma redução no desempenho do classificador, quando utilizada à técnica de filtro. Porém, a Árvore de Decisão gerada foi menor. Esta menor quantidade de nodos permite uma melhor visualização das regras obtidas. Algumas regras associaram internet banda larga, laboratório de ciências, auditório na escola e ensino privado, como os fatores mais influentes na conclusão do Ensino Fundamental.

Guércio *et al.* (2014) propõem a criação de um modelo que analisa o comportamento dos alunos em diferentes períodos de tempo, permitindo assim, avaliar se um aluno possui tendência a aprovação ou reprovação antes do término da disciplina. Foi analisado o desempenho em disciplinas ministradas em um curso de graduação. Os dados foram extraídos do AVA *Moddle* e do Sistema de Gerenciamento Acadêmico (SGA) utilizado na instituição de ensino. Foram utilizados os algoritmos *Random Forest*, *Random Tree* e *J48* presentes na ferramenta *Weka*. Segundo os autores, a acurácia média encontrada foi de 73% na classificação e foi possível observar a possibilidade de transformar os dados armazenados na base de dados da plataforma *Moodle* em conhecimento, gerando regras muito úteis para o apoio à tomada de decisões.

Dos Santos *et al.* (2014) identificam primeiro as disciplinas que mais influenciam na evasão do curso. Depois, utilizam essas disciplinas para predição de desempenho dos alunos ao longo das semanas. Foi analisado o desempenho em disciplinas ofertadas em cursos de graduação a distância. Os dados foram obtidos no AVA *Moodle* e no SGA da instituição de ensino. Os algoritmos utilizados foram *SimpleCart*, *J48* e o *ADTree* presentes na ferramenta *Weka*. A partir dos experimentos realizados, os autores constataram 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 que 80%.

Com o objetivo geral de investigar a possibilidade de geração de inferências relativas ao desempenho futuro de estudantes, Gottardo (2013) utiliza técnicas de mineração de dados utilizando atributos disponíveis em uma base de dados do AVA *Moodle*. Em seu quinto experimento, o autor utiliza o conceito de séries temporais obtendo taxas de acurácia próximas a 72%, mesmo em etapas iniciais de realização do curso. Segundo o autor, estas informações poderiam ser úteis para o desenvolvimento de ações envolvendo estudantes da turma em andamento e não apenas de turmas futuras.

3. Materiais e Métodos

Com os objetivos de predizer o desempenho acadêmico durante a realização de um curso, e de extrair conhecimento potencialmente útil para gestores, este trabalho aplica o processo de KDD, empregando a tarefa de classificação em bases de dados extraídas de um AVA *Moodle* e de um SGA utilizado por uma IFES.

Os dados utilizados neste trabalho foram retirados do curso de graduação em Administração, ofertado na modalidade à distância por uma IFES. O processo de escolha da disciplina para representar o curso foi baseado nos seguintes critérios:

- Disciplina específica do curso ofertada no primeiro ano da graduação;
- Número de estudantes matriculados;

- Quantidade de recursos utilizados;
- Quantidade de interações dos alunos com o AVA *Moodle.*

Este trabalho utiliza atributos que representam o comportamento do estudante na plataforma, ou seja, a forma como o ele interage com as atividades e recursos disponibilizados. A partir da tabela de *log* do AVA *Moodle,* foram extraídos dados quantitativos sobre o comportamento e participação dos estudantes durante a realização do curso, enquanto do SGA foram retiradas as notas dos estudantes. A Tabela 1 apresenta os atributos extraídos.

Tabela 1 – Relação dos atributos utilizados

Grupo	Atributos	Descrição
Tarefas	up_tarefas	Envio de atividades
	all_tarefas	Acesso a todas as atividades disponíveis no curso
	page_tarefas	Acesso à página de determinada atividade
Blog	ad_blog	Adição de *blog*
	acesso_blog	Acesso ao *blog*
Chat	acesso_chat	Acesso à sala do *chat*
	msg_chat	Mensagens enviadas no *chat*
	rel_chat	Acesso aos *chats* realizados
Questionários Múltipla Escolha	resp_quest_nov	Responde questionário novamente
	resp_quest	Responde questionário
	acesso_quest	Acesso ao questionário
	acesso_quest_all	Acesso a todos os questionários
Curso	acesso_curso	Acesso ao curso
Materiais	acesso_material_all	Acesso a todas as pastas de conteúdo
	acesso_material	Acesso à pasta de conteúdos
Fórum	del_post_forum	Quantidade de postagens de uma discussão apagados
	ad_discussao_forum	Adição de uma discussão no fórum
	insc_forum_all	Se inscrever para receber atualizações de todos os fóruns
	enc_forum	Encerrar inscrição para receber atualizações de um fórum
	acesso_forum	Acesso à página principal do fórum de discussão
	del_discussao_forum	Quantidade de discussões de um fórum apagadas
	enc_forum_all	Encerrar inscrição para receber atualizações de todos os fóruns
	upd_post_forum	Atualização de post na discussão do fórum
	acesso_discussao_forum	Acesso à determinada discussão dentro de um Fórum
	ad_post_forum	Adição de post na discussão do fórum
	acesso_foruns	Acesso a todos os fóruns disponibilizados no curso
	insc_forum	Se inscrever para receber atualizações de um fórum
Informações	acesso_inf	Acesso à página de informações do curso (ementa, bibliografia,...)
Recursos	acesso_lista_recursos	Acesso a todos os materiais disponibilizados no curso
	acesso_recurso	Visualização de apostilas ou vídeos disponibilizados
Vídeos	acesso_video	Acesso a vídeos em sites externos
	acesso_video_all	Acesso à listagem de todos os vídeos em sites externos
Usuários	upd_perfil	Atualização do perfil
	acesso_user_all	Acesso à página com todos os usuários do curso
	acesso_user	Acesso a determinado usuário
Wiki	comentario_wiki	Acesso a determinado comentário da *wiki*
	acesso_dif_wiki	Acesso a comparação entre versões da *wiki*
	hist._wiki	Histórico de modificações na *wiki*
	acesso_wiki	Acesso à *wiki*
	restaurar_wiki	Restaurar versão da *wiki*
	ad_item_wiki	Adição de um novo item na *wiki*
	comentarios_wiki	Acesso a todos os comentários da *wiki*
	acesso_mapa_wiki	Acesso ao mapa da *wiki*
	upd_wiki	Edição de um item da *wiki*
	Classe	Classe a ser prevista pelos algoritmos de classificação

A base do Curso foi dividida em 2 classes (Aprovado e Reprovado) através do processo conhecido como discretização (estudantes foram alocados em determinada classe de acordo com a sua nota final). A classe Aprovado compreende os estudantes que obtiveram notas entre 60 e 100, enquanto na classe Reprovado estão os alunos com notas entre 0 e 59 pontos. A distribuição completa do conjunto de dados obtido pode ser observada na Tabela 2.

Tabela 2 – Conjunto de dados obtido e suas respectivas classes

Conjunto	Nº de total de estudantes	Nº de estudantes na Classe Aprovado	Nº de estudantes na Classe Reprovado
Administração	248 (100%)	120 (48,4%)	128 (52,6%)

A ferramenta utilizada nos experimentos foi o *Weka.* Ela reúne um conjunto de algoritmos de Aprendizado de Máquina para tarefas de Mineração de Dados [Bouckaert *et al*., 2015]. Optou-se por escolher e testar os seguintes algoritmos de classificação presentes nesta ferramenta, *AdaBoost* [Freund e Schapire, 1996], *BayesNet* [Heckerman, 1995], IBk [Aha *et al*., 1991], J48 [Quinlan, 1993], *Random Forest* [Breiman, 2001], JRip [Cohen, 1995], *Multilayer Perceptron* [Rumelhart *et al*., 1986] e SVM [Platt, 1998], todos utilizados, ou indicados em alguns trabalhos pesquisados tais como Kotsiantis *et al*., 2007; Wu *et al*., 2008; Hall *et al*., 2011. Para avaliar o desempenho dos algoritmos de classificação, este trabalho utiliza a métrica Acurácia (Ac.), que possui valores entre 0 e 100%.

Para a realização da tarefa de classificação é necessário que o algoritmo utilizado na construção do modelo de aprendizado seja treinado. Para realizar este treinamento, e posterior teste, é necessário que o conjunto de dados seja dividido em dois conjuntos distintos, treinamento e teste. Com esta finalidade, este trabalho utiliza o método (bem difundido) de validação cruzada de 10 *folds* (10-*fold cross validation*). [Han *et al.*, 2012; Mitchel, 1997].

Dentre as técnicas de Seleção de Atributos disponíveis, foi escolhida a abordagem Cápsula (*Wrapper*), porque dentre os trabalhos pesquisados, nenhum fez uso deste método em conjuntos de dados retirados de cursos de graduação, na modalidade à distância, ofertados em AVA's. Nesta abordagem o algoritmo de indução é utilizado para avaliar, e consequentemente selecionar o melhor subconjunto de atributos [Kohavi *et al*., 1996].

4. Experimentos

Neste experimento, foi aplicado o conceito de Séries Temporais para propor um modelo de acompanhamento de estudantes durante a realização de um curso a distância. Na IFES, a oferta de disciplinas nos cursos à distância é realizada de forma semestral. Durante o semestre são realizadas duas avaliações presenciais, sendo este o procedimento padrão para todos os cursos de graduação ofertados.

O período de oferta da disciplina foi dividido em 8 períodos. Como pode ser observado na Figura 1, os períodos possuem duração de aproximadamente 15 dias, com exceção do primeiro que possui 25 dias. Devido aos ajustes de matrículas e a ambientação do estudante na plataforma, optou-se pelo maior número de dias neste primeiro período.

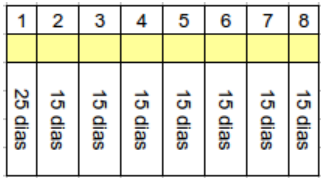

Figura 1 – Divisão dos períodos da série temporal

A primeira avaliação presencial acontece no final do período 3, e a partir do período 4, esta nota parcial é inserida como um atributo adicional ao conjunto de dados original, conforme a Figura 2.

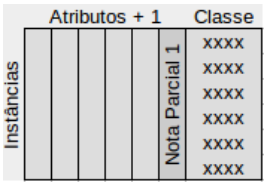

Figura 2 – Conjunto de dados original mais nota da primeira avaliação

A segunda avaliação presencial acontece ao final do sétimo período e no período 8. Esta segunda nota parcial é inserida como um segundo atributo adicional ao conjunto de dados, como pode ser observado na Figura 3.

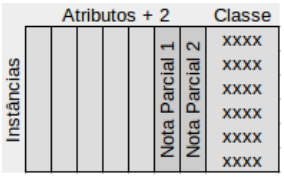

Figura 3 – Conjunto de dados original mais as notas da primeira e segunda avaliação

Como a oferta da disciplina foi dividida em 8 períodos de tempo, foram gerados 8 conjuntos de dados, onde o conjunto 1 contém as interações dos estudantes com a plataforma nos 25 primeiros dias da oferta. O conjunto de dados 2 contém as interações dos 25 dias do período 1 mais os 15 dias do período 2, e assim sucessivamente até o oitavo conjunto de dados.

O objetivo da divisão do conjunto de dados em séries temporais é que os responsáveis pela turma (professores, tutores) tenham o *feedback* do progresso do estudante e possam acompanhá-los durante a realização do curso. Nesta proposta de divisão, os responsáveis terão 3 *feedbacks* antes da primeira avaliação e mais 4 entre as duas avaliações presenciais. Dessa forma, os objetivos deste estudo são:

- Analisar o desempenho do modelo de previsão em séries temporais no processo de acompanhamento do estudante durante a realização da disciplina;
- Analisar o desempenho do modelo de previsão em séries temporais ao inserir as notas parciais das avaliações no processo de acompanhamento do estudante durante a realização da disciplina;
- Analisar o desempenho do modelo de previsão em séries temporais ao inserir as notas parciais das avaliações com a utilização de seleção de atributos;
- Analisar a viabilidade de avaliar o desempenho de estudantes em estágios precoces da disciplina.

A Tabela 3 mostra os resultados dos algoritmos aplicados às séries temporais utilizando o conjunto de dados original. Tem-se o melhor desempenho ao longo da série com algoritmo *BayesNet* que obteve uma média de 68,1% de acurácia nos 8 conjuntos temporais.

Pode-se observar um baixo desempenho dos algoritmos classificadores no início da série, e conforme progride-se ao oitavo período, tem-se uma melhora no desempenho

dos classificadores. Este fato ocorre devido ao aumento do volume de dados gerado pelas interações dos estudantes com a plataforma no decorrer do tempo, melhorando o treinamento e consequente aprendizado dos classificadores.

Tabela 3 – Resultados para a classificação do conjunto original em séries temporais

Algoritmo	1	2	3	4	5	6	7	8	Ac. Média
AB	64,5%	61,3%	62,1%	62,9%	**68,5%**	62,9%	65,7%	67,3%	64,4%
BN	67,3%	66,1%	**66,5%**	**66,5%**	67,7%	**69,0%**	69,8%	71,8%	**68,1%**
IBk	60,1%	55,2%	63,3%	61,7%	62,1%	63,7%	66,1%	66,9%	62,4%
J48	59,7%	64,1%	61,7%	59,3%	65,7%	63,7%	64,5%	63,3%	62,8%
JRip	64,1%	**66,5%**	65,7%	65,7%	62,9%	61,3%	**70,2%**	66,1%	65,3%
MP	59,7%	61,3%	60,5%	60,1%	62,5%	62,9%	65,3%	62,5%	61,9%
RF	**68,5%**	60,9%	63,3%	60,9%	60,9%	61,3%	66,5%	**72,2%**	64,3%
SVM	65,3%	64,5%	63,3%	65,3%	64,9%	64,1%	66,5%	67,3%	65,2%
Ac. Média	63,7%	62,5%	63,3%	62,8%	64,4%	63,6%	66,8%	67,2%	---

A Tabela 4 apresenta os resultados da classificação do conjunto original de dados com as notas das duas avaliações presenciais. Pode-se observar um aumento considerável no desempenho de todos os algoritmos classificadores testados ao inserir a primeira nota parcial a partir do período 4. Ao inserir a primeira nota parcial, o algoritmo *AdaBoost* obteve o melhor desempenho dentre os classificadores utilizados, passando de 62,1% de acurácia no período 3 para 73,8% no período 4. O melhor desempenho médio para os oito períodos foi do algoritmo JRip com 71,7% de acurácia.

Tabela 4 – Resultados para a classificação do conjunto original mais notas parciais em séries temporais

Algoritmo	1	2	3	4	5	6	7	8	Ac. Média
AB	64,5%	61,3%	62,1%	**73,8%**	74,6%	74,6%	75,8%	79,0%	70,7%
BN	67,3%	66,1%	**66,5%**	69,4%	69,4%	69,4%	72,2%	78,2%	69,8%
IBk	60,1%	55,2%	63,3%	69,8%	71,0%	68,5%	69,9%	75,4%	66,7%
J48	59,7%	64,1%	61,7%	69,4%	69,0%	69,0%	70,6%	75,0%	67,3%
JRip	64,1%	**66,5%**	65,7%	73,0%	**75,0%**	**75,0%**	**77,4%**	77,0%	**71,7%**
MP	59,7%	61,3%	60,5%	71,4%	70,2%	70,2%	70,2%	79,4%	67,9%
RF	**68,5%**	60,9%	63,3%	68,5%	71,0%	71,0%	71,0%	**80,2%**	69,3%
SVM	65,3%	64,5%	63,3%	72,2%	73,0%	73,0%	72,6%	**80,2%**	70,5%
Ac. Média	63,7%	62,5%	63,3%	70,9%	71,7%	71,3%	72,5%	78,1%	---

Na Tabela 5 pode-se observar os resultados para a classificação do conjunto de dados com Seleção de Atributos pelo método Cápsula mais as notas das duas avaliações presenciais. Com a utilização da Seleção de Atributos o desempenho melhorou de forma

significativa, e todos os algoritmos classificadores elevaram as suas taxas de acurácia em relação aos resultados para o conjunto de dados original. Destaca-se a melhoria de desempenho do *Multilayer Perceptron* que obteve a melhor média para os oito períodos, 73,6% de acurácia. Com a inserção da nota da primeira avaliação presencial, a acurácia obtida foi de 76,2%, aumento de 9,7% em relação aos testes realizados no conjunto de dados original.

Tabela 5 – Resultados para a classificação do conjunto usando a técnica de seleção de atributos pelo método cápsula mais notas parciais em séries temporais

Algoritmo	1	2	3	4	5	6	7	8	Ac. Média
AB	64,1%	65,7%	67,7%	71,8%	71,0%	73,0%	**75,0%**	77,8%	70,8%
BN	65,7%	66,5%	66,5%	68,5%	70,2%	72,6%	73,4%	79,4%	70,4%
IBk	58,1%	56,5%	55,2%	65,3%	64,9%	66,1%	66,9%	78,6%	64,0%
J48	60,9%	58,9%	64,1%	**76,2%**	73,0%	75,0%	72,6%	78,6%	69,9%
JRip	**68,1%**	63,7%	**71,0%**	73,4%	74,2%	**75,8%**	**75,0%**	80,2%	72,7%
MP	67,7%	68,1%	70,6%	74,2%	**75,0%**	74,6%	73,8%	**84,7%**	**73,6%**
RF	63,3%	61,3%	64,5%	75,0%	71,0%	73,4%	**75,0%**	80,6%	70,5%
SVM	66,1%	**69,4%**	67,7%	73,8%	73,0%	73,4%	73,4%	79,8%	72,1%
Ac. Média	64,3%	63,8%	65,9%	72,3%	71,5%	73,0%	73,1%	80,0%	---

4.1. Análise

A Tabela 6 apresenta a comparação dos resultados médios em cada período de tempo obtidos pelas três abordagens apresentadas neste experimento. A coluna Média representa o aumento médio da acurácia obtido pelos classificadores utilizando o conjunto de dados original mais as notas parciais, e a técnica de seleção de atributos mais as notas parciais, em relação ao conjunto de dados original.

Tabela 6 – Comparação dos resultados das três abordagens propostas no experimento com séries temporais

Bases	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Média**
Conjunto original	63,7%	62,5%	63,3%	62,8%	64,4%	63,6%	66,8%	67,2%	---
Conjunto original mais notas parciais	63,7%	62,5%	63,3%	70,9%	71,7%	71,3%	72,5%	78,1%	7,9%
Seleção de Atributos mais nota parcial	64,3%	63,8%	65,9%	72,3%	71,5%	73,0%	73,1%	80,0%	9,0%

O aumento médio informado na Tabela 6 é referente ao intervalo entre os períodos 4 e 8, pois é a partir do quarto período que são inseridas as notas parciais. É possível observar que no conjunto original e no conjunto original mais notas parciais, os três primeiros resultados são iguais, isso porque os dados são os mesmos. Para o

conjunto que utiliza a seleção de atributos mais as notas parciais, houve aumento da acurácia nos três primeiros períodos.

Utilizando o conceito de Séries Temporais, as inferências relativas ao desempenho dos estudantes obtiveram 68,1% de acurácia logo no primeiro corte, e após a inserção da nota da primeira avaliação presencial, chegou-se a 76,2% de acurácia no conjunto que utiliza a abordagem Cápsula de Seleção de Atributos. Tais resultados demonstram a viabilidade das técnicas já em períodos iniciais do curso.

Os resultados da aplicação da abordagem Cápsula de Seleção de Atributos, até então não utilizada em trabalhos correlatos, mostrou um aumento significativo no desempenho dos classificadores na realização de inferências sobre o desempenho acadêmico dos estudantes em relação ao conjunto de dados original, chegando a um aumento médio de 9% de acurácia.

5. Conclusões e Trabalhos Futuros

Este trabalho analisou e extraiu conhecimento a partir dos bancos de dados do AVA *Moodle* e do SGA de uma IFES. A possibilidade de acompanhar o desempenho de um estudante durante a realização de cursos EAD é de grande importância para professores, tutores e gestores, que podem ajustar suas ações pedagógicas para evitar a reprovação destes estudantes.

Com a utilização do conceito de séries temporais, buscou-se a identificação precoce de estudantes com maior probabilidade de baixo desempenho. Foram propostas três abordagens, sendo a abordagem que utiliza a Seleção de Atributos pelo método Cápsula com a inserção das notas das avaliações presenciais, a mais recomendada para o problema. Utilizando o conceito de séries temporais e realizando os cortes temporais conforme a Figura 1, os gestores podem acompanhar o progresso dos estudantes durante a realização do curso. São três *feedbacks* antes da primeira avaliação presencial e mais 4 até a segunda avaliação. Através destes *feedbacks*, os gestores das turmas podem realizar intervenções pedagógicas com o objetivo de melhorar o desempenho dos estudantes.

Como trabalhos futuros, pretende-se realizar novos experimentos utilizando a mesma base de dados para aplicar a Seleção de Atributos pelo método filtro utilizando os algoritmos *Cfs Subset Eval, Chi Squared, Attribute Eval, Info Gain Attribute Eval, Gain Ratio Attribute Eval, Relief FAttribute Eval,* utilizados em alguns trabalhos relacionados como Ramaswami e Bhaskaran (2009) e Marquez-Vera, Romero e Ventura (2011).

Referências

Aha, D., Kibler, D. & Albert, M. (1991). Instance-Based Learning Algorithms. MachineLearning,6(1):37–66.

Baker, R. S. J., Isotani, S. & de Carvalho, A. M. J. B. (2011). Mineração de dados educacionais: Oportunidades para o brasil. Revista Brasileira de Informática na Educação, 19(2), pp. 3-13.

Breiman, L. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

Bouckaert, R., et al., 2015, WEKA Manual for Version 3-6-4. 2010. Disponível em: http://ufpr.dl.sourceforge.net/project/weka/documentation/3.7.x/WekaManual-3-7-13.pdf. Acesso em: 16 jun. 2016.

Cohen, W. (1995). Fast Effective Rule Induction. In Proceedings of ICML-95, 115-123.

dos Santos, R. N., de Alburqueque Siebra, C., & Oliveira, E. S. (2014). Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão. In: Anais dos Workshops do III Congresso Brasileiro de Informática na Educação.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), pp. 37-54.

Ferreira, G. (2015). Investigação acerca dos fatores determinantes para a conclusão do Ensino Fundamental utilizando Mineração de Dados Educacionais no Censo Escolar da Educação Básica do INEP 2014. In: Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação.

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In L. Saitta (Ed.), Proceedings of the Thirteenth International Conference on Machine Learning (pp. 148–156). Bari, Italy. San Francisco: Morgan Kaufmann.

Gottardo, E. (2013). Estimativa de desempenho acadêmico de estudantes em um AVA utilizando técnicas de mineração de dados (Dissertação Mestrado). Universidade Tecnológica Federal do Paraná.

Guércio, H., Marques, P., Ströele, V., Pereira, C. K., & Barrere, E. (2014). *Análise do Desempenho Estudantil na Educação a Distância Aplicando Técnicas de Mineração de Dados.* In: Anais dos Workshops do III Congresso Brasileiro de Informática na Educação.

Hall, M., Witten, I. & Frank, E. (2011). Data mining: Practical machine learning tools and techniques. Kaufmann, Burlington.

Hand, D. J., Mannila, H. & Smyth, P. (2001). Principles of data mining. MIT press.

Heckerman, D.(1995). A Tutorial on Learning With Bayesian Networks. Microsoft Research, Page 57. number MSR-TR-95-06, March.

Kohavi, R., Sommerfield, D. & Odugherty, J. (1996) Data Mining Using MLC++, A Machine Learning Library in C++. In Proceedings of the 8th International Conference on Tools with Artificial Intelligence (ICTAI '96). IEEE Computer Society, Washington, DC, USA, 234-245, 1996.

Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. (2007). Supervised machine learning: A review of classification techniques. 3-24.

Marquez-Vera, C., Romero, C., Ventura, S. (2011). Predicting school failure using data mining. In: EDM. p. 271–276.

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

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, & A. Smola (Eds.), Advances in kernel methods: Support vector learning (pp. 185–209). Cambridge, MA: MIT Press.

Quinlan, J.R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.

Ramaswami, M., Bhaskaran, R. (2009) A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by backpropagating errors. Nature, 323:533–53.

Wu, X. et al. (2008). Top 10 algorithms in data mining. Knowledge and information systems, v. 14, n. 1, p. 1-37.