2013 SEGeT SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA
Gestão e Tecnologia para a Competitividade
23.24.25 de Outubro de 2013

A PREDIÇÃO DA EVASÃO DE ESTUDANTES DE GRADUAÇÃO COMO RECURSO DE APOIO FORNECIDO POR UM ASSISTENTE INTELIGENTE

Átila Pires dos Santos
atila.santos@ifb.edu.br
IFB, UCB

Sandra Isaelle Figueiredo dos Santos
sandraisaelle@gmail.com
IFB

Vandor Roberto Vilardi Rissoli
vandor@ucb.br
UCB

Resumo:A evasão no ensino superior tem sido objeto frequente de estudo. Pesquisas de combate a este nível de evasão escolar estão acontecendo com apoio de sistemas computacionais, sendo neste trabalho realizada análises significativas que incluem a tecnologia de um Sistema Tutor Inteligente, denominado SAE, e alguns métodos de mineração de dados. Em sua proposta, almeja-se combinar estas duas tecnologias para o processamento adequado ao seu contexto de estudo (evasão no ensino superior), tendo como objetivo potencializar seus resultados a fim de combatê-la. Foi utilizada a base de dados do SAE, centrada em seus aprendizes, com algoritmos de mineração de dados de clustering e de classificação. Como resultado, os alunos foram classificados em cinco grupos, do menos propenso à evasão para o mais propenso ou já evadido. Dessa forma, tal combinação se mostrou muito propensa ao apoio coerente ao combate da evasão, além de propiciar novas possibilidades de intervenção em tempo real a este problema educacional.

Palavras Chave: Evasão Escolar - Mineração de Dados - STI - Cluster - TI na Educação

1. INTRODUÇÃO

A educação no nível superior tem enfrentado alguns entraves que prejudicam seu sucesso, destacando-se, entre eles, a evasão discente. Diante desta realidade, seus gestores procuram tomar decisões estratégicas visando sanar este problema no âmbito educacional. No entanto, a eficácia destas decisões tem níveis variados de sucesso, sendo geralmente relacionados à quantidade e qualidade das informações disponíveis em tempo adequado ao apoio destas decisões. Andriola et al (2006, p. 368) questiona o grau de sucesso que as gestões educacionais vêm obtendo no contexto da evasão, uma vez que os "dados acerca da evasão discente no ensino superior são pouco explorados, acarretando, consequentemente, diminuta compreensão do fenômeno e de suas causas".

Pesquisas têm sido realizadas buscando resolver estes entraves ao sucesso escolar, onde recursos tecnológicos estão sendo empregados como ferramentas de apoio às atividades educacionais. Têm-se destacado, entre estes recursos, os softwares educacionais classificados como Sistemas Tutores Inteligentes (STI), voltados ao apoio do processo de ensino e ao atendimento das necessidades individuais de seus estudantes.

Uma implementação deste tipo de software, utilizada por algumas universidades brasileiras, é denominada SAE (Sistema de Apoio Educacional), STI que emprega a tecnologia ITA (*Intelligent Teaching Assistant*) para fornecer orientação adequada a cada um de seus usuários-estudantes a partir de suas interações com o próprio sistema, além dos vários "pontos de vista" de outros agentes humanos (monitores e docentes) colaboradores com o sucesso deste intrincado processo de ensino e aprendizagem.

Este tipo de STI, reconhecido como Assistente Virtual de Ensino Inteligente, ainda coleta informações que podem ser usadas em decisões estratégicas quanto às posturas didático-pedagógicas que devem ser assumidas por estes outros agentes humanos em relação às características individuais de cada aprendiz. Por meio da coleta e armazenamento desses dados significativos a modelagem dos aspectos cognitivos de cada aprendiz é possível agir, interventivamente, sobre cada estudante que esteja interessado em sua própria aprendizagem.

O SAE, no entanto, não realiza uma análise mais formal acerca da evasão escolar, abordando este tema apenas indiretamente, combatendo algumas de suas causas. Por outro lado, foram realizadas abordagens diretas ao tema da evasão através de sistemas de mineração de dados (*data mining*). O objetivo dessa mineração é a busca de padrões existentes em grandes quantidades de dados armazenados, de maneira persistente (banco de dados), por sistemas computacionais (WITTEN; FRANK, 2005). Através da mineração de dados disponíveis sobre um universo, é possível predizer comportamentos de componentes deste universo que compartilham características em comum, por exemplo, nos aspectos relacionados à evasão discente.

As abordagens de combate à evasão discente apoiada pela mineração de dados, geralmente, fazem uso das informações provenientes das bases de dados de registros acadêmicos armazenados sobre os estudantes. O que é então fornecido por estas bases à mineração para análise da evasão é um conjunto de dados cadastrais, que pouco variam ao longo do tempo (como, por exemplo, o sexo, o endereço, etc.), ou uma coleção de informações acerca das notas finais e controle de frequência destes alunos matriculados em uma disciplina, que são renovados a cada período letivo (normalmente semestral no ensino superior brasileiro).

Estas abordagens, no entanto, ao invés de detectar um problema que pode acontecer, estão muitas vezes detectando um problema que já ocorreu, uma vez que estas se baseiam em dados que são atualizados uma vez a cada período letivo. Caso, ao longo de um período

letivo, o contexto de um aluno se altere, de maneira a incentivá-lo a evadir de seu curso, a mineração de dados pode ter detectado esta situação tarde demais e o estudante, possivelmente, já ter evadido.

A abordagem ao problema da evasão discente pode ter resultados mais promissores, caso possa ser acompanhada de dados que sejam atualizados com maior frequência. A proposta deste artigo almeja combinar estes dois tipos de tecnologias, STI e Mineração de dados, de maneira adequada ao processamento conjunto em um mesmo contexto do ensino superior (a evasão), com o objetivo de potencializar seus resultados no combate à evasão.

Este artigo está dividido em 5 seções, sendo na seção 2 apresentada sua plataforma teórica sobre a mineração de dados no contexto da evasão e nos princípios da tecnologia dos Sistemas Tutores Inteligentes como Assistentes Inteligentes. A seção 3 explica a metodologia utilizada e sua forma de acompanhamento mais realista da situação de aprendizagem de cada estudante, com apoio de um Assistente Virtual de Ensino Inteligente. Na seção 4 são relatados os principais resultados alcançados com este experimento, enquanto a seção 5 apresenta as principais considerações sobre os resultados obtidos nesta pesquisa inicial. As referências bibliográficas são relacionadas ao final deste artigo.

2. REFERENCIAL TEÓRICO

O primeiro passo para predizer e combater a evasão é buscar por um conceito do que é a evasão. Dore e Lüscher (2011, p. 775) discorrem sobre o problema da definição do termo evasão, tendo este sido associado a diversas situações, como "a retenção e repetência do aluno na escola, a saída do aluno da instituição, a saída do aluno do sistema de ensino, a não conclusão de um determinado nível de ensino, o abandono da escola e posterior retorno" e também "àqueles indivíduos que nunca ingressaram em um determinado nível de ensino, especialmente na educação compulsória, e ao estudante que concluiu um determinado nível de ensino, mas se comporta como um *dropout*", ou seja, como um aluno evadido. Desta maneira, torna-se difícil desassociar a evasão do sucesso ou insucesso escolar.

Gaioso (2005, p. 9) define a evasão como sendo a "interrupção no ciclo de estudo". Na pesquisa elaborada por este autor (p. 38), considerou-se o aluno como evadido quando o mesmo "deixou o curso por qualquer motivo que não seja a obtenção da titulação", listando como alternativas à conclusão do curso como sendo: abandono, ou seja, "a matrícula não foi efetuada no curso dentro do prazo estabelecido"; transferência interna ou mudança de curso; transferência externa; matrícula em curso de outra instituição via aprovação em processo seletivo; e "desistência, re-opção ou jubilamento".

Uma das primeiras tentativas de explicar o fenômeno da evasão através de um modelo teórico foi feito por Tinto (1975). Ele argumenta que a decisão de um estudante pela permanência ou evasão se baseia na integração deste estudante ao ambiente do curso que ele ingressa. Esta integração, por sua vez, influencia e é influenciada pelas intenções, objetivos e compromissos que o mesmo possui. Dekker et al (2009) e Hatos e Suta (2011) argumentam sobre a aparente predominância do modelo de Tinto no campo de estudos acerca da evasão estudantil. Entretanto, os autores Hatos e Suta criticam a capacidade preditiva deste modelo para muitos casos e Andriola et al (2006, p. 367) ressalta que este modelo "tal como foi proposto, não se aplica, em sua totalidade, à realidade brasileira".

2.1. A MINERAÇÃO DE DADOS

Dekker et al (2009, p. 42) considera o uso de técnicas de mineração de dados para predição da evasão discente como sendo "relativamente nova". Durante o levantamento de trabalhos relevantes sobre este tema foi confirmada tal afirmação, pois nenhuma publicação

foi encontrada antes do ano 2000. Além disso, grande parte da bibliografia utilizada por estes trabalhos, como referencial teórico, é composta por conteúdos que não tratam diretamente o problema da evasão discente, tendo como objeto de estudo o sucesso ou insucesso escolar. Dentre os trabalhos que efetivamente abordam a evasão discente, com emprego de técnicas de mineração de dados, merecem destaque neste artigo Superby et al (2006), Dekker et al (2009), Bayer et al (2012), além de Campello e Lins (2008), Cobbe et al (2011) e Manhães et al (2011, 2012) no contexto brasileiro.

A abordagem utilizada por estes artigos é semelhante, variando principalmente no quesito de quais atributos foram trabalhados e qual ou quais algoritmos de mineração de dados foram utilizados. Os atributos dos estudantes mais frequentemente encontradas nesses estudos foram: status/notas do aluno em disciplinas já cursadas (incluindo aqui o total de reprovações que o aluno já teve e sua frequência), sexo/estado civil, idade, profissão, renda familiar e o endereço residencial.

A escolha destas variáveis é justificada por Cobbe et al (2011) pelo levantamento feito por Gaioso (2005), que apresenta como algumas das principais causas da evasão do nível superior: a necessidade de trabalhar / horário de trabalho incompatível com o de estudo, problemas financeiros, casamento / nascimento de filhos, desconhecimento da metodologia do curso escolhido, deficiência da educação básica e reprovações sucessivas.

Gaioso (2005) ainda aponta como causas da evasão a falta de orientação vocacional / profissional e imaturidade, a ausência de perspectivas de trabalho, a falta de laços afetivos na universidade, busca de desafio a si mesmo (ou seja, buscar ser admitido em um curso sem a intenção de cursá-lo, apenas para demonstrar sua capacidade em ser admitido para aquele curso), a herança profissional (matricular-se em um curso escolhido pela família do discente, não por ele mesmo), a ausência de um referencial na família, mudança de endereço e concorrência entre as instituições de ensino superior.

Os fatores que resultam em evasão levantados por este autor podem ser divididos em dois grupos: os fatores externos às Instituições de Ensino Superior (IES) e os fatores internos as IES. Estão no primeiro grupo os fatores que fazem parte do histórico do estudante (como a deficiência da educação básica) e da vida do aprendiz fora da IES (como trabalho e família). No segundo grupo, estão os fatores que são resultados da relação do estudante com seu curso, como o desconhecimento da metodologia do curso escolhido e as reprovações sucessivas.

Vale ressaltar que há ainda outras variáveis, presentes em apenas alguns estudos: nota do aluno no processo seletivo de admissão na IES, a estrutura familiar do aprendiz e o nível de escolaridade do pai e/ou mãe do estudante. Superby et al (2006) vai além, trazendo como variável em seu estudo até mesmo o fato do estudante ser fumante ou não.

Quanto à escolha do algoritmo de mineração de dados, a maioria dos trabalhos pesquisados utiliza mais do que um tipo e os compara e/ou os agrega em um comitê. Dekker et al (2009) conclui que, possivelmente, os algoritmos de classificação por árvore de decisão sejam as melhores escolhas e que os algoritmos de *clustering* podem auxiliar na criação das classes e categorização dos estudantes dentro destas classes, como feito por Campello e Lins (2008).

2.2. O SISTEMA DE APOIO EDUCACIONAL

No Sistema de Apoio Educacional (SAE) cada estudante interage com os recursos fornecidos por este ITA (YACEF, 2002), que utiliza a Lógica Fuzzy em seus processos de inferência para modelar os aspectos cognitivos de cada aprendiz. Este assistente inteligente (SAE) serve de tutor virtual aos aprendizes durante os seus momentos mais propícios de dedicação ao estudo e a fixação do conteúdo a ser assimilado, por meio da realização de

exercícios e tarefas, interação com os monitores e participação de atividades colaborativas com os demais estudantes de sua turma, além de seu próprio professor. No ambiente virtual do SAE cada aprendiz pode acompanhar seu próprio esforço e desempenho obtido sobre cada conteúdo de aprendizagem, sendo indicado por este sistema a possível necessidade de maior atenção e mais tempo no estudo sobre os tópicos específicos de um conteúdo em que estão sendo detectadas deficiências momentâneas de assimilação, podendo estas comprometerem a aprendizagem de cada estudante.

As análises e as variáveis linguísticas acompanhadas pelo SAE, sobre a realidade de cada aprendiz, são explicitadas por Rissoli e Santos (2011), sendo seu principal foco a averiguação dos termos linguísticos mais coerentes as variáveis: *participação, esforço* e *desempenho* do aprendiz. A primeira variável, *participação*, infere "a participação de cada aprendiz nas atividades interativas propostas pelo docente", que "poderão acontecer por meio de fóruns ou chats (bate-papo) envolvendo assuntos relacionados aos conceitos pertinentes a cada conteúdo". A variável *esforço* averigua "o número de exercícios resolvidos e a quantidade de visitas que cada estudante efetuou na monitoria estudantil" por tópico ou conceito abrangido por cada conteúdo, enquanto que o *desempenho* "envolve o resultado obtido na solução da quantidade de exercícios apurados pela variável *esforço*" (RISSOLI; SANTOS, 2011, p. 7-9).

Estes autores subdividem os aspectos relacionados a qualidade obtida na variável *desempenho* em: tipo de questão (verdadeira ou falsa, múltipla escolha, escolhas múltiplas, lacuna e aberta ou dissertativa); nível de dificuldade (fácil, médio e difícil) e categoria da questão (revisão, fixação e avaliativa). Cada registro de uma questão no SAE pode assumir um único valor em cada um destes três aspectos responsáveis pela apuração fuzzy que qualifica o resultado obtido. Por exemplo, uma questão do tipo aberta, que seja avaliativa e possua nível de dificuldade difícil, poderia representar um valor expressivo à inferência fuzzy para esta variável linguística (*desempenho*) apurada pelo SAE. É importante ressaltar que cada tipo de questão, nível de dificuldade e categoria trabalham uma habilidade e capacidade de reflexão condizente com a assimilação do conteúdo por cada aprendiz, sendo relevante a combinação mais coerente entre estes recursos de análise do SAE, conforme a expectativa de aprendizagem desejada em cada conceito que compõe um conteúdo (ou disciplina).

Além de orientar pedagogicamente os aprendizes, este ITA ainda coleta dados que são relevantes à assistência adequada para as possíveis mudanças de estratégias didáticas na atuação mais coerente do docente e dos monitores estudantis. Através destes dados torna-se possível interceder junto a um aprendiz que esteja tendo dificuldades com o conteúdo estudado. Rissoli (2007, p.30) ressalta que "estas dificuldades podem contribuir ainda mais com a falta de motivação do estudante" e que isto pode promover o "acumulo de conteúdo a ser estudado, timidez em expressar suas dúvidas em sala de aula, principalmente pela dificuldade em formular suas questões, além da não participação efetiva nos trabalhos elaborados em grupo", e que, desta maneira, poderia contribuir com "a evasão escolar.". Sendo assim, através do SAE e de informações sobre o contexto do aprendiz fora da IES, é possível predizer, com maior segurança, um quadro de possível evasão em tempo real, possibilitando uma ação de maior intervenção docente e do próprio ITA junto a cada estudante que o utilize como recurso de apoio educacional.

3. METODOLOGIA

A pesquisa elaborada neste artigo classifica-se, em relação ao seu objetivo, como explicativa, e quanto a sua natureza como quantitativa. A pesquisa explicativa tem "como preocupação central identificar os fatores que determinam ou que contribuem para a ocorrência dos fenômenos" (GIL, 1991). A natureza dos dados estudados permite que seja

realizada uma análise quantitativa, ou seja, a realização de averiguações sobre dados numéricos coletados (MARTINS; THEÓPHILO, 2009), neste caso através de técnicas de mineração de dados.

Para este estudo foi adotado o conceito de evasão proposto por Gaioso (2005, p. 38), anteriormente já apresentado, sendo este quando o estudante "deixou o curso por qualquer motivo que não seja a obtenção da titulação". Infere-se a partir desta definição que o estudante evadido, ou em processo de evasão, realiza uma série de abandonos, em cinco etapas: *1*) abandono da disciplina que está cursando; *2*) abandono do semestre ou módulo; *3*) desistência do curso; *4*) deixar a instituição (IES); *5*) abandono do nível superior como um todo.

Depreende-se que a segunda etapa tem como pré-requisito a primeira (seja ela ao longo de um período letivo, seja ela entre um período letivo e outro). Da mesma forma, para alcançar a terceira etapa, ele deverá ter realizado a segunda, que implica ter realizado a primeira. O mesmo vale para a quarta e quinta etapas. Desta forma, um estudante que alcance as últimas etapas de abandono deverá ter passado antes pela primeira etapa.

Diante destas constatações, o presente estudo busca agir interventivamente sobre o problema da evasão através da predição da primeira etapa do abandono, sendo este um pré-requisito para o alcance das outras etapas. Para tanto, foram utilizados os dados de aprendizes extraídos do SAE em períodos letivos anteriores. Como este ITA está restrito apenas ao contexto de sala de aula, sem conter dados pertinentes de um sistema de registro acadêmico, este estudo limita-se à predição do abandono de disciplinas durante o transcorrer do período letivo.

A base de dados dos estudantes disponibilizada pelo SAE para este estudo continha 242 instâncias ao todo, cada uma representando um aluno do segundo semestre de 2012. Estes alunos cursavam o Bacharelado em Sistemas de Informação (BSI), o Bacharelado em Ciência da Computação (BCC) ou um curso livre, sendo identificados, respectivamente, neste trabalho como alunos de BSI (118 no total), de BCC (105) e 19 do curso livre. As disciplinas analisadas foram: Algoritmo (33 alunos de BSI e 46 alunos de BCC), Laboratório de Programação 1 (40 alunos de BCC), Laboratório de Programação 2 (19 alunos de BCC), Linguagem e Técnicas de Programação 1 (45 alunos de BSI), Linguagem e Técnicas de Programação 2 (40 alunos de BSI) e Tecnologias Inteligentes no Apoio à Educação (19 alunos do curso livre que envolvia docentes como estudantes).

Durante a fase de pré-processamento foram selecionadas, para cada uma destas instâncias, as seguintes variáveis: *i*) total de questões obrigatórias respondidas pelo aluno, *ii*) total de questões obrigatórias respondidas corretamente pelo aluno, *iii*) total de questões avulsas respondidas pelo aluno, *iv*) total de questões avulsas respondidas corretamente pelo aluno, *v*) número de acessos realizados ao SAE pelo aluno, *vi*) nome da disciplina, *vii*) nome do professor, *viii*) idade do aluno e *ix*) sexo do aluno. As variáveis de *i* a *v* foram utilizadas para a divisão do universo de alunos em *clusters* e todas as variáveis foram utilizadas para a etapa de classificação.

Para as duas etapas seguintes, de *clustering* e classificação, foi usado o software de mineração de dados Weka (*Waikato Environment for Knowledge Analysis*), que é uma ferramenta genérica e livre para a mineração, suportando diferentes abordagens de aprendizagem de máquina (Washio et al, 2007). Como observado por Dekker et al (2009) e executado por Campello e Lins (2008), fez-se uso de *clusters* para agrupar instâncias similares, com o objetivo de criar categorias para classificar a situação de cada aluno. Dois algoritmos diferentes foram utilizados neste experimento, ambos usando suas configurações padrão exceto pelo valor do parâmetro *k*, configurado para criar cinco *clusters*. Estes dois algoritmos foram o *SimpleKMeans* e o EM. Segundo Witten e Frank (2005, p. 137), *k-means*

é a "técnica de *clustering* clássica", sendo um método "simples e efetivo". Já o algoritmo EM, sigla para *Expectation-Maximization*, é descrito pelo mesmo autor (p. 265) como sendo capaz de "convergir para um máximo de maneira garantida", ainda que este seja um máximo local e não global.

Para a etapa de classificação o algoritmo utilizado foi *Random Committee*, cujo nome completo é *Random Tree based Committee Learning*. Ele é descrito por Washio et al (2007) como sendo um comitê que usa por base um número pré-determinado de *Random Trees*. A opção por este algoritmo se baseou na observação feita por Dekker et al (2009) acerca da possível melhor escolha de algoritmos de classificação para o problema da evasão.

A escolha pelo uso de algoritmos de *cluster* se deve a necessidade de classificar as instâncias extraídas do banco de dados do SAE. Por um lado, esta necessidade acontece pelo fato de não haver uma classificação nativa ao SAE acerca da evasão discente, sendo essencial, portanto, que este dado seja fornecido por outra fonte. Por outro lado, independente da fonte consultada, esta classificação estará dividida em apenas duas classes (evadido ou não evadido), não sendo possível assim observar as nuances dentro de cada classe.

Os algoritmos de classificação foram trazidos neste trabalho com uma alternativa aos algoritmos de *cluster*. Uma vez que eles são algoritmos de aprendizagem de máquina supervisionada (enquanto os algoritmos de *cluster* trabalham com aprendizagem não-supervisionada), há a necessidade que as classes estejam pré-definidas antes de sua utilização (WITTEN; FRANK, 2005). Desta forma, eles podem ser utilizados apenas depois que as instâncias já foram associadas aos *clusters*, usados aqui como classes. Após esta primeira etapa, no entanto, a aprendizagem supervisionada torna-se uma alternativa viável e interessante a esta análise.

4. RESULTADOS

Os dois universos de cinco *clusters* criados pelos dois algoritmos (*SimpleKMeans* e EM) demonstraram ser consistentes entre si, pois ainda que estes não tenham produzido resultados idênticos, eles foram capazes de categorizar os estudantes em dois contextos bastante distintos: os alunos não evadidos e os alunos evadidos ou em risco de evadir. Em apenas dez instâncias (ou seja, 4% da amostra) os dois algoritmos discordaram acerca desta fronteira básica.

Os *clusters* que ambos os algoritmos criaram parecem representar linearmente os estudantes menos propensos aos mais propensos ou já evadidos. Esta relação também pode ser interpretada como a representação dos alunos mais propensos a obter sucesso na disciplina para os menos propensos. Desta forma, denominou-se cada *cluster* como: ***a***) Alunos Muito Participativos, ***b***) Alunos Participativos, ***c***) Alunos Pouco Participativos, ***d***) Alunos Propensos a Evadir e ***e***) Alunos Evadidos.

A principal diferença entre as classificações de cada estudante dentro de um *cluster*, feitas por um dos dois algoritmos, se deu no dimensionamento final que cada *cluster* obteve, em especial os *clusters* ***d*** (propensos) e ***e*** (evadidos). *SimpleKMeans* dimensionou seus *clusters* da seguinte maneira: 13 alunos para ***a***, 34 alunos para ***b***, 46 alunos para ***c***, 101 alunos para ***d*** e 48 para ***e***. O algoritmo EM dimensionou seus *clusters* da seguinte forma: 10 alunos para ***a***, 28 alunos para ***b***, 51 alunos para ***c***, 26 alunos para ***d*** e 127 para ***e***.

Dos alunos classificados pelo *SimpleKMeans* como evadidos, apenas três deles não foram considerados como evadidos pelo EM, mas como propensos a evadir. Destes 45 alunos classificados por ambos os algoritmos como evadidos, apenas sete foram incorretamente classificados; ainda assim, apenas 3 deles obtiveram aprovação na disciplina.

Os 3 alunos que foram classificados como evadidos apenas pelo *SimpleKMeans* de fato são alunos evadidos. Dos 82 alunos que foram classificados como evadidos apenas pelo EM, apenas 20 foram incorretamente classificados, 18 dos quais obtiveram aprovação na disciplina.

A relação entre estas cinco classes (***a***, ***b***, ***c***, ***d***, ***e***) e o sexo dos estudantes indica que ambos (masculino e feminino) estão representados em todas as classes, mesmo que o sexo feminino represente apenas 10% do todo (25 instâncias). Ambos os algoritmos classificaram uma aluna como da classe ***a*** e outra da classe ***b***, enquanto seis estavam na classe ***c***. Para o *SimpleKMeans* há sete alunas da classe ***d*** e dez alunas da classe ***e***, enquanto para o EM há três alunas da classe ***d*** e quatorze alunas da classe ***e***.

Embora haja alunos mais jovens representados nas cinco classes, os dois algoritmos concordam que neste universo, quanto mais velho é o aluno, maior sua propensão a evadir. Para o algoritmo *SimpleKMeans*, não havia alunos com idade igual ou superior a 25 anos nas classes ***a*** ou ***b*** e, para o algoritmo EM, seria 30 ou superior. Dos 31 alunos acima da faixa de 30 anos ou mais, apenas sete (para o *SimpleKMeans*) ou cinco (para o EM) estavam fora das classes ***d*** e ***e***.

A relação entre as sete disciplinas analisadas e as cinco classes está representada nas duas tabelas a seguir (Tabela 1 com *SimpleKMeans* e Tabela 2 com EM). As disciplinas Algoritmo (BCC), Laboratório de Programação 1 e Laboratório de Programação 2 foram bem representadas, em ambos os algoritmos, nas cinco classes. Por outro lado, as outras quatro disciplinas tiveram resultados preocupantes, em especial Linguagem e Técnicas de Programação 2, que possuía alunos somente nas classes ***d*** e ***e***.

Tabela 1: Relação entre classes e disciplinas para o *SimpleKMeans*.

SimpleKMeans	Classe ***a***	Classe ***b***	Classe ***c***	Classe ***d***	Classe ***e***
Algoritmo (BSI)	0	0	4	20	9
Algoritmo (BCC)	6	20	12	6	2
Laboratório de Programação 1	6	10	7	15	2
Laboratório de Programação 2	1	4	7	5	2
Linguagem e Técnicas de Programação 1	0	0	10	23	12
Linguagem e Técnicas de Programação 2	0	0	0	30	10
Tecnologias Inteligentes no Apoio à Educação	0	0	6	2	11
Total	**13**	**34**	**46**	**101**	**48**

Tabela 2: Relação entre classes e disciplinas para o EM

EM	Classe ***a***	Classe ***b***	Classe ***c***	Classe ***d***	Classe ***e***
Algoritmo (BSI)	0	0	2	0	31
Algoritmo (BCC)	5	14	18	5	4
Laboratório de Programação 1	3	13	5	3	16
Laboratório de Programação 2	1	1	10	5	2
Linguagem e Técnicas de Programação 1	1	0	11	6	27
Linguagem e Técnicas de Programação 2	0	0	0	4	36
Tecnologias Inteligentes no Apoio à Educação	0	0	5	3	11
Total	**10**	**28**	**51**	**26**	**127**

A relação entre os resultados e os professores também demonstrou uma situação preocupante. Quatro professores, denominados neste estudo como *A*, *B*, *C* e *D*, ministraram as disciplinas anteriormente listadas. Para o algoritmo *SimpleKMeans* haviam 13 alunos do professor *A* na classe ***a***, 23 alunos na classe ***b***, 24 alunos na classe ***c***, 19 alunos na classe ***d*** e 18 alunos na classe ***e***. Para o professor *B*, haviam 11 alunos na classe ***b***, 22 alunos na classe ***c***, 23 alunos na classe ***d*** e 3 alunos na classe ***e***. Para o professor *C*, haviam 49 alunos na classe ***d*** e 18 alunos na classe ***e***, enquanto que para o professor *D* haviam 10 alunos na classe ***d*** e 9 alunos na classe ***e***.

Para o algoritmo EM haviam 10 alunos do professor *A* na classe ***a***, 22 alunos na classe ***b***, 29 alunos na classe ***c***, 20 alunos na classe ***d*** e 16 alunos na classe ***e***. Para o professor *B*, haviam 6 alunos na classe ***b***, 22 alunos na classe ***c*** e 31 alunos na classe ***e***. Para o professor *C*, haviam 6 alunos na classe ***d*** e 61 alunos na classe ***e***, enquanto que para o professor *D* haviam 19 alunos na classe ***e***.

Assim, enquanto os professores *B* e, principalmente, *A* obtiveram resultados bastante equilibrados de seus alunos, os professores *C* e *D* possuíam apenas estudantes nas classes ***d*** e ***e***. No entanto, isso evidencia apenas que estes professores não incentivam seus alunos a fazer uso do ambiente SAE durante suas aulas, causando a falsa impressão de que estes aprendizes haviam evadido. Por exemplo, dos 19 alunos do professor *D*, apenas dois efetivamente evadiram e apenas seis não obtiveram sucesso nesta disciplina. Esta visão também justifica o desequilíbrio percebido entre as classes e as disciplinas, uma vez que o professor *C* lecionou as disciplinas Linguagem e Técnicas de Programação 1 e 2 e o professor *D* lecionou Algoritmo (BSI).

Foi utilizada na última etapa deste experimento, envolvendo algoritmos de classificação, o algoritmo *Random Committee*. Este algoritmo foi configurado para realizar 50 iterações, como sugerido por Washio et al (2007). Utilizou-se tanto o método de *cross-validation* quanto *percentage split* neste experimento, usando *10-fold* para o primeiro e 33,33% para o segundo. A escolha do primeiro se justifica por ele possibilitar uma melhor estimativa da margem de erro de classificação (WITTEN; FRANK, 2005), enquanto o segundo procura simular uma situação mais próxima da realidade. Estes dois testes, mencionados anteriormente, foram realizados tanto para as classificações criadas pelo *SimpleKMeans* e pelo EM, resultando em quatro testes distintos.

Para a classificação pelo *SimpleKMeans*, os resultados obtidos foram de 91.7355% em *10 fold* e 90.7407% para o *split* de 33,33%. Para a classificação pelo EM, os resultados obtidos foram de 95.8678% em *10 fold* e 88.8889% para o *split* de 33,33%. Estes resultados são condizentes com os resultados obtidos por outros trabalhos semelhantes, como Dekker et al (2009), Cobbe et al (2011) e Manhães et al (2011, 2012).

5. CONCLUSÃO

O uso de técnicas de mineração (*data-mining*) nos dados provenientes do SAE sobre seus aprendizes mostrou-se promissor neste estudo. Por um lado, os dados fornecidos pelo SAE possibilitaram predições sólidas aos algoritmos de mineração. Por outro lado, a criação de *clusters* permitiu a análise dos dados contidos no SAE de modo que não lhe era possível fazer antes. Foi observado ainda que existe uma forte influência da disciplina cursada e do professor que a ministra sobre a evasão discente percebida pela mineração de dados. Esta influência acontece não apenas na dimensão didático-pedagógica, facilitando ou dificultando ao estudante a conclusão de uma disciplina, mas também na relação entre o professor e o SAE, tornando os dados dos estudantes armazenados por este assistente virtual (SAE) mais ou menos ricos. Este aspecto se apresentou como uma fragilidade às análises deste experimento,

mas propiciou a apuração inicial sobre os modelos educacionais adotados pelos docentes que utilizam o SAE como recurso de apoio educacional. Diante destas análises foi possível inferir que a postura dos docentes *C* e *D* se mantém mais tradicionais, ou seja, trabalham o ensino centrado no professor, enquanto que os docentes *A* e *B* labutam a maior autonomia em seus aprendizes e trabalham o ensino-aprendizagem mais centrado na aprendizagem de cada estudante.

Apesar desta fragilidade, também foi observado que a combinação destas tecnologias (STI/ITA e Mineração de dados) possui ainda potenciais inexplorados, sendo inúmeras suas possibilidades futuras. Como o SAE analisa vários dados, provenientes de diferentes perfis de usuários (aluno, monitor, professor), e infere, continuamente, novas informações sobre seus estudantes, este processamento de classificação do potencial de evasão discente também poderia acontecer em tempo real. Desta forma, seria possível acompanhar a evolução de cada aprendiz ao longo de seu período letivo, sendo cada um ainda assistido sobre a situação de transição entre as possíveis classes de evasão. Este novo recurso de apoio, fornecido pelo SAE, se constituiria em uma nova variável linguística a ser incorporada a sua base de conhecimento, o que aumentaria suas possibilidades de assistência e, consequentemente, ampliaria o potencial de acompanhamento para a tomada de decisões mais estratégicas ao êxito educacional.

6. REFERÊNCIAS

ANDRIOLA, W. B; ANDRIOLA, C. G; MOURA, C. P. Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da Universidade Federal do Ceará (UFC). Ensaio: Aval. Pol. Públ. Educ., Rio de Janeiro, v.14, n.52, p.365-382, 2006.

BAYER, J; BYDZOVSKA, **H; GERYK, J; OBSIVAC, T; POPELINSKY, L.** Predicting drop-out from social behaviour of students. In: Anais 5th International Conference on Educational Data Mining- EDM 2012, Chania, Grécia, 2012.

CAMPELLO, A. V. C; LINS, L. N. Metodologia de análise e tratamento da evasão e retenção em cursos de graduação instituições federais de ensino superior. In: Anais XXVIII Encontro Nacional de Engenharia de Produção, Rio de Janeiro, 2008.

COBBE, P. R; BALANIUK, R; PRADO, H. A; GUADAGNIN, R. V; FERNEDA, E. Predicting evasion candidates in higher education institutions. Model and Data Engineering, Brasília, v. 6918, p. 143-151, 2011.

DEKKER, G, PECHENIZKIY, M; VLEESHOUWERS, J. Predicting Students Drop Out: A Case Study. In Anais Proceedings of the International Conference on Educational Data Mining, Córdoba, Espanha, 2009, p. 41-50.

DORE, R; LÜSCHER, A. Z. Permanência e evasão na educação técnica de nível médio em Minas Gerais. Cad. Pesqui., São Paulo, v. 41, n. 144, 2011 .

GAIOSO, N. P. L. Evasão discente na educação superior: a perspectiva dos dirigentes e dos alunos. Brasília: UCB, 2005, 99 P.

GIL, A. C. Como elaborar projetos de pesquisa. São Paulo. Atlas. 1991

HATOS, A; SUTA, G. Student persistence in higher education. A literature review. Higher Education Research and Development (HERD), Oradae, v.1 , 2011.

MANHÃES, L. M. B; CRUZ, S. M. S; COSTA, R. J. M; ZAVALETA, J; ZIMBRÃO, G. Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. In: Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo, 2012.

MANHÃES, L. M. B; CRUZ, S. M. S; COSTA, R. J. M; ZAVALETA, J; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: Anais do XXII SBIE - XVII WIE, Aracaju, 2011.

MARTINS, G. A; THEÓPHILO, C R. Metodologia da Investigação Científica para Ciências Sociais Aplicadas. 2 ed. São Paulo: Atlas, 2009.

RISSOLI, V. R. V; SANTOS, G. A. Um Assistente Inteligente Fuzzy no Acompanhamento da Aprendizagem Significativa. In: Congresso da Sociedade Brasileira de Computação, 2011, Natal. Anais do Congresso da Sociedade Brasileira de Computação. Porto Alegre: SBC, 2011. v. 31.

RISSOLI, V. R. V. Uma proposta metodológica de acompanhamento personalizado para Aprendizagem Significativa apoiada por um Assistente Virtual de Ensino Inteligente. Porto Alegre: PGIE, 2007. 224 p. Tese (Doutorado) Universidade Federal do Rio Grande do Sul, 2007.

SUPERBY, J; VANDAMME, J.-P; MESKENS, N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In Proc. of the Workshop on Educational Data Mining at ITS'06, p. 37-44, 2006.

TINTO, V. Dropout from higher education: a theoretical synthesis of recent research. Review of Educational Research, New York, n. 45, p. 89-125, 1975.

WASHIO, T; SHINNOU, Y; YADA, K; MOTODA, H; OKADA, T. Analysis on a Relation Between Enterprise Profit and Financial State by Using Data Mining Techniques. Springer v. 4384, p.306-316, 2007.

WITTEN, I. H; FRANK, E. Data mining: practical machine learning tools and techniques. 2 ed. San Francisco: Elsevier, 2005.

YACEF, K. Intelligent Teaching Assistant Systems. In: International Conference on Computers in Education, 2002. New Zeland. Proceedings International Conference on Computers in Education. New Zeland: IEEE, 2002. v. 1, p. 136-140.