Evasão de estudantes universitários: diagnóstico a partir de dados acadêmicos e socieconômicos

Túlio Albuquerque Pascoal[1], **Daniel Miranda de Brito**[1], **Leandro Paiva Andrade**[1], **Thaís Gaudencio do Rêgo**[1]

[1]Universidade Federal da Paraíba (UFPB)
Caixa Postal 5.115 - 58.051-970 – João Pessoa – PB – Brasil

{tuliopascoal,britmb,leandropiox,gaudenciothais}@gmail.com

Abstract. *Educators and managers of higher education institutions have been concerned about the high dropout rates of their students. Thus, means of early diagnosis of students prone to evasion are desired, in order to avoid that to happen. In this paper, it is proposed an approach to assist institutions in making decisions to combat this phenomenon, using academic and socioeconomic data collected from the students. The viability of the strategy was studied based on records of Computer Science students at UFPB. The results obtained were significant, with accuracy rates of more than 85% in the classification of students.*

Resumo. *Educadores e gestores das instituições de ensino superior têm se preocupado com os altos índices de evasão de seus estudantes. Dessa forma, meios para o diagnóstico precoce de estudantes propensos à evasão são desejados, de forma a evitá-la. Neste trabalho, propõe-se uma abordagem para auxiliar as instituições nas tomadas de decisão para combate desse fenômeno, a partir do uso de dados acadêmicos e socioeconômicos dos estudantes. A viabilidade da estratégia foi estudada com base em registros de alunos do curso de Ciência da Computação da UFPB. Os resultados obtidos foram relevantes, com taxas de acerto de mais de 85% na classificação dos estudantes.*

1. Introdução

A evasão no ensino superior é um problema internacional e impacta diretamente a eficiência dos sistemas educacionais [Silva Filho et al. 2007]. Caracterizada como a interrupção dos ciclos de estudo, gera prejuízos significativos em vários aspectos, tais como: sociais, acadêmicos e econômicos [Aparecida et al. 2011] [Baggi e Lopes 2010]. No setor público, as perdas provocadas por estudantes que evadem de seus cursos representam recursos públicos investidos na educação sem o retorno apropriado para a sociedade, enquanto no setor privado acarreta importantes perdas de receita. Em ambos os casos a evasão é fonte de inocupação de pessoal e de equipamentos [Silva Filho et al. 2007].

[Barroso e Falcão 2004], em seu estudo com alunos do curso de Física da Universidade Federal do Rio de Janeiro (UFRJ) discutem a evasão segundo três enfoques distintos: econômico, vocacional e institucional. O primeiro diz respeito à impossibilidade de manutenção do vínculo do aluno por questões socioeconômicas; o segundo se refere à evasão em termos de uma escolha inadequada por parte do aluno, tornando-o desmotivado com a área de graduação; e, por fim, a evasão institucional tem relação com o fracasso nas disciplinas iniciais do curso, seja por inadequação dos métodos de estudo,

DOI: 10.5753/cbie.sbie.2016.926

ou por deficiências prévias de conteúdo, impedindo que a aprendizagem ocorra de forma efetiva. Na direção da ideia do primeiro enfoque, a desistência dos alunos também pode ocorrer em virtude de assuntos particulares à vida do estudante, como nos casos em que ingressam no mercado de trabalho e, com isso, experimentam dificuldades em conciliar os horários de estudos com os do emprego [Gisi 2006] [de Souza 2008].

Dentre as áreas que apresentam maiores taxas de evasão, Ciências, Matemática e Computação evidenciam-se, com taxas bem acima da média nacional [Silva Filho et al. 2007]. Em relação aos cursos de Computação, é comum, nas universidades brasileiras, que seus estudantes apresentem dificuldades no aprendizado de disciplinas de programação e lógica, levando a um alto índice de reprovação nas disciplinas específicas da área e, em alguns casos, a desistência do curso [Prietch e Pazeto 2010]. [Silva Filho et al. 2007] constataram que a evasão ocorre com frequência de duas a três vezes maior nos primeiros períodos do curso, portanto, estratégias que visem solucionar o problema da evasão em estágios iniciais do curso tendem a ser adequadas.

Tendo em vista a importância do controle do fenômeno da evasão para a manutenção dos sistemas de ensino de nível superior, apresenta-se neste trabalho, com auxílio de técnicas de mineração de dados educacionais (EDM, do inglês *educational data mining*), uma abordagem para a identificação de estudantes propensos a evasão, com base nos seus dados socioeconômicos e de desempenho, de modo que educadores possam tomar decisões apoiados no sistema.

A organização do restante do texto é a seguinte: na Seção 2 apresentam-se e discutem-se os trabalhos relacionados encontrados na literatura; na Seção 3 explica-se e detalha-se a metodologia da abordagem proposta e as métricas de avaliação são definidas; a discussão e análise dos resultados obtidos encontram-se na Seção 4 e finalmente, na Seção 5 tem-se a conclusão e possíveis trabalhos futuros.

2. Trabalhos Relacionados

Define-se EDM como uma disciplina emergente, preocupada com a criação de estratégias para investigar os diversos tipos de informações originárias de ambientes educacionais, de forma que se possa ter uma melhor compreensão dos estudantes e do sistema educacional em que eles aprendem [Baker et al. 2010]. Assim, a sua utilização permite um melhor entendimento dos dados originados pelo ambiente educacional, através da geração de conhecimentos adicionais, como descoberta de padrões e correlações implícitas nos dados. Portanto, a EDM pode ser utilizada para auxiliar as tomadas de decisão e planejamento dos sistemas de educação, beneficiando todos os afetados direta ou indiretamente (instituição, coordenadores, professores, estudantes, etc.) [Romero e Ventura 2007]. Na Figura 1, é apresentada uma visão geral de como a EDM se relaciona e pode beneficiar os sistemas educacionais.

O interesse nas pesquisas sobre a utilização de métodos de mineração de dados no campo educacional é crescente, principalmente naquelas que tratam do tema da evasão no ensino superior. Nos últimos anos, essas pesquisas tornaram-se evidentes, sejam as desenvolvidas no âmbito das instituições de ensino nacionais, sejam no das instituições internacionais. Ademais, os estudos são desenvolvidas tanto em ambientes presenciais, quanto em ambientes a distância [Romero e Ventura 2010]. [Vitelli et al. 2011] realizaram um estudo sobre evasão de estudantes na Universidade do Vale do Rio dos Sinos

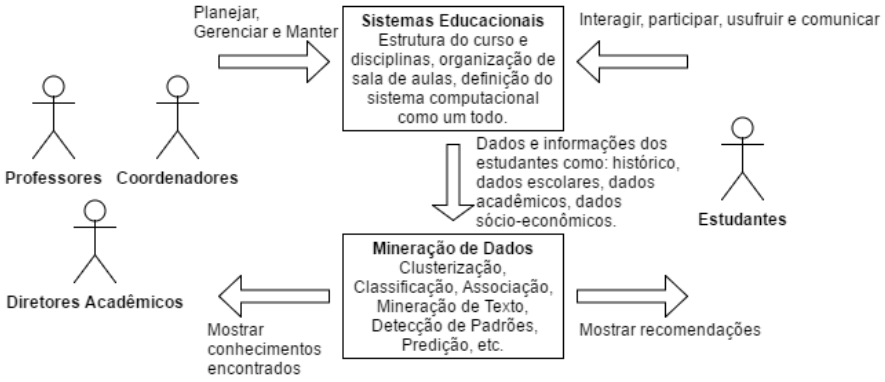

Figura 1. Ciclo de aplicação da Mineração de Dados Educacionais. Adaptado de: [Romero e Ventura 2007] pg. 136

utilizando dados socioeconômicos, desempenho no vestibular, média de desempenho nas atividades, entre outros. Foram aplicados métodos de regressão logística para indicar os fatores que ocasionavam a evasão e quantificar sua relevância. Os resultados revelaram que os fatores mais impactantes estão relacionados ao desempenho acadêmico e ao tempo de curso.

[de Brito et al. 2014] examinam a probabilidade de predição de desempenho de alunos do primeiro semestre do curso de Ciência da Computação da UFPB, utilizando as notas de ingresso na instituição como base. Com o uso de classificadores, os autores alcançaram taxas de acerto de até 75%. Eles argumentam que a identificação do desempenho de alunos nas disciplinas do primeiro período é útil para combater a evasão produzida pelo fracasso nas disciplinas iniciais. Ainda utilizando o ambiente da UFPB, [Pascoal et al. 2015] propuseram um método para previsão de desempenho de alunos de Computação em disciplinas de programação. Foram utilizados quatro algoritmos de aprendizagem de máquina distintos, os quais produziram taxas de acerto sempre acima de 70%, chegando a 84,7% no melhor caso, com o algoritmo *Random Forest*.

No estudo de caso de [Dekker et al. 2009], descrevem-se os resultados obtidos na previsão de evasão de estudantes do curso de Engenharia Elétrica, da Universidade de Eindhover. Os autores consideram como critério de predição a educação pré-universitária e as notas dos estudantes no primeiro semestre do curso. Utilizando a ferramenta Weka, os autores conseguiram resultados com taxas de acerto entre 75% e 80% na identificação de estudantes com risco de evasão. Em outra pesquisa, [Osmanbegović e Suljić 2012] aplicaram três algoritmos supervisionados de mineração de dados: *Naive* Bayes, rede neural e árvore de decisão, para prever o sucesso ou o fracasso dos alunos em um curso e avaliaram o desempenho preditivo dos métodos de aprendizagem com base em sua precisão da previsão, facilidade de aprendizagem, e as características de fácil utilização. Os resultados indicaram que o classificador *Naive* Bayes supera, por sua precisão da previsão, os outros métodos considerados.

[Silva et al. 2014] realizaram um estudo em EDM utilizando regras de associação, a fim de encontrar padrões de regras entre os resultados das provas e os questionários socioeconômicos do Exame Nacional do Ensino Médio (ENEM). Os autores concluíram que

uma baixa renda familiar, escolaridade dos pais restrita ao nível primário e quantidade elevada de pessoas que residem com os alunos são fatores que reduzem o seu desempenho na prova. Considerando a problemática da evasão, [de Oliveira Júnior et al. 2014] analisaram a correlação entre o empréstimo de livros na biblioteca e a permanência do aluno na instituição. Também foram utilizadas informações de desempenho no ENEM (forma de ingresso na instituição) e informações sociais de gênero e idade. Os autores observaram uma correlação próxima a 80% entre os atributos de entrada e a situação final do aluno no curso.

3. Metodologia

3.1. Método proposto

Com intuito de otimizar os sistemas de ensino, objetiva-se neste trabalho disponibilizar uma metodologia para a identificação precoce de estudantes propensos à evasão, de modo a facilitar ações dos educadores que visem reduzir os índices de evasão e permitir que os alunos permaneçam nos cursos. Estudantes são classificados como propensos à evasão ou possíveis concluintes, segundo informações socioeconômicas e de desempenho acadêmico, e a viabilidade das predições é estudada através de testes na ferramenta Weka. Os detalhes da abordagem são apresentados em seguida.

No método proposto, utiliza-se o algoritmo de classificação *Naive* Bayes. Optou-se por sua escolha, em virtude da sua ampla utilização em diversos trabalhos da área [Bhardwaj e Pal 2012] [Baradwaj e Pal 2012] [Xenos 2004], os quais apresentam bons resultados de classificação. O algoritmo é baseado no teorema de Bayes, que classifica uma instância não rotulada com base na maior probabilidade dela estar associada aos seus atributos de entrada, isto é, da hipótese mais provável. A classificação é fundamentada na distribuição de probabilidade de uma classe, dada a probabilidade condicional de seus atributos, assumindo independência entre eles (*Naive*) [Ang et al. 2016] [Faceli et al. 2011]. O método pode ser resumido a partir da Equação 1:

$$h = argmax_i P(y_i) \prod_{j=1}^{d} P(\mathbf{x}_j | y_i) \quad (1)$$

O classificador *Naive* Bayes é considerado descritivo e preditivo, sendo assim, apto e recomendado para realizar classificações preditivas. Além de não necessitar de um grande conjunto de treinamento, sua aplicação é fácil, pois só necessita de uma única passagem pelo conjunto de treinamento e requer somente cálculos simples, como de média, variância e probabilidade [Bhardwaj e Pal 2012]. Além disso, o classificador é robusto, mesmo com a presença de ruído e atributos irrelevantes, e apresenta um bom desempenho em uma grande variedade de domínios [Faceli et al. 2011]. A eficiência e simplicidade do método igualmente são fatores de destaque [Osmanbegović e Suljić 2012].

Na Figura 2, uma visão geral da abordagem proposta é apresentada, em termos do seu fluxo de informações e processos. Primeiramente, um conjunto de dados é analisado e classificado pelo algoritmo *Naive* Bayes, tendo como resultado um modelo (hipótese) criado a partir da natureza dos dados e gerado pelo algoritmo de classificação. Uma vez em posse desse método e dos dados acadêmicos e socioeconômicos de um estudante, os interessados podem utilizar a hipótese aprendida para produzir um rótulo de classe para

aquele aluno, ou seja, se ele tem risco de evasão no curso ou não. Dada a qualidade dessa previsão, tomadas de decisões e ações para evitar a evasão serão facilitadas (caso sejam necessárias) para os envolvidos no sistema educacional. Uma vez que, por exemplo, o quanto antes um coordenador de curso souber da chance de um de seus alunos evadir, medidas mais efetivas podem ser tomadas para evitar o seu abandono. A última etapa do método, mas não menos importante, é a incorporação dos dados do aluno (juntamente com a classificação resultante do modelo) ao conjunto de dados do sistema, a fim de manter, após uma nova classificação, um conjunto de dados sempre atualizado, de forma a produzir resultados cada vez mais confiáveis, consistentes e condizentes com o cenário.

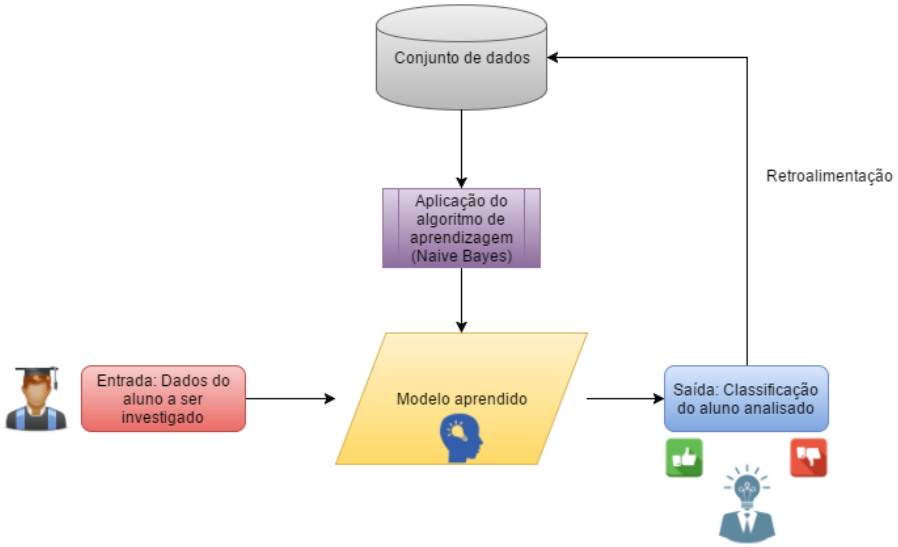

Figura 2. Fluxograma do método proposto

3.2. Descrição dos dados

Para estudar a viabilidade da abordagem proposta, utilizaram-se informações referentes a estudantes do curso de Ciência da Computação da Universidade Federal da Paraíba (UFPB). As informações foram cedidas pela Superintendência de Tecnologia da Informação (STI) da Universidade e consistem de registros de estudantes do ano de 2001 a 2013, fornecidos em dois arquivos distintos. O primeiro, contendo os dados referentes às disciplinas cursadas na Universidade e as notas de ingresso (aqui referidas como dados acadêmicos), e a segunda com os dados socioeconômicos dos estudantes. As tabelas foram unidas e o resultado pré-processado, com intuito de eliminar as instâncias com dados ausentes e inconsistentes. Restaram 241 instâncias para análise, sendo 40 da classe **concluinte** (alunos que concluíram o curso) e 201 da classe **evadido** (alunos que se evadiram do curso). Para a tarefa de classificação, consideraram-se todos os 22 atributos com valores não ausentes resultantes da união das tabelas, separados em duas modalidades: acadêmicos e socioeconômicos, que podem ser vistos nas Tabelas 1 e 2, respectivamente.

3.3. Métricas de avaliação

A fim de avaliar a qualidade das predições de estudantes, utilizou-se a ferramenta Weka (versão 3.6) [Hall et al. 2009], que provê um método simples para avaliação de algoritmos aprendizagem de máquina [Witten e Frank 2011]. Utilizou-se para a classificação o

Tabela 1. Atributos acadêmicos levados em consideração pela classificação

Atributo	Descrição	Tipo
nota_calculo_1	Nota obtida na disciplina	Numérico (0 a 10)
nota_calculo_vet	Nota obtida na disciplina	Numérico (0 a 10)
nota_fisica_1	Nota obtida na disciplina	Numérico (0 a 10)
nota_intro_prog	Nota obtida na disciplina	Numérico 0 a 10)
nota_intro_comp	Nota obtida na disciplina	Numérico (0 a 10)
escola_ens_fund	Categoria de ensino frequentado pelo aluno no ensino fundamental	Nominal (parte em escola particular e parte em escola pública, tendo ficado mais tempo em escola particular; parte em escola pública e parte em escola particular, tendo ficado mais tempo em escola pública; somente em escola particular; somente em escola pública)
escola_ens_medio	Categoria de ensino frequentado pelo aluno no ensino médio	Nominal (parte em escola particular e parte em escola pública, tendo ficado mais tempo em escola particular; parte em escola pública e parte em escola particular, tendo ficado mais tempo em escola pública; somente em escola particular; somente em escola pública)
turno_ens_medio	Turno frequentado no ensino médio	Nominal (integral (dois turnos); parte diurno e parte noturno, predominando o diurno; parte diurno e parte noturno, predominando o noturno; somente diurno; somente noturno)
mediageral	Média obtida pelo aluno no exame de admissão na Universidade	Numérico (0 a 1000)

algoritmo *Naive* Bayes. A divisão entre o conjunto de treinamento e testes foi feita pela Validação Cruzada (do inglês *Cross- Validation*[1]) de 10 *folds* (padrão do Weka).

Como métricas de avaliação de desempenho da abordagem proposta, usou-se a acurácia (número de instâncias classificadas corretamente); matriz de confusão e as métricas dela derivadas (verdadeiros positivos, VP; verdadeiros negativos, VN; falsos positivos, FP; e falsos negativos, FN); e a área sob a curva ROC (AUC, do inglês *area under the curve*). A matriz de confusão mostra o desempenho de um algoritmo de classificação a partir da relação de qual(is) registro(s) do conjunto de dados foram classificados corretamente. As métricas de VP e VN indicam, respectivamente, classificações corretas dos estudantes que abandonaram e dos estudantes concluíram o curso. Já as taxas de FN e FP correspondem, respectivamente, a estudantes evadidos classificados como concluintes e estudantes concluintes classificados como evadidos. Previamente à análise das classificações, o ganho de informação de cada atributo foi estudado com a finalidade de identificar os mais importantes para a decisão do algoritmo [Witten e Frank 2011].

4. Discussão dos Resultados

Nesta seção, apresentam-se os resultados obtidos a partir da análise do conjunto de dados dos alunos, em dois experimentos no Weka. O primeiro analisou a capacidade preditiva de cada atributo para a classificação final do aluno, enquanto o segundo avaliou a eficiência do modelo gerado para a predição de novas instâncias. Inicialmente foram realizados testes somente utilizando os atributos socioeconômicos dos alunos, sem levar em consideração os dados acadêmicos. Adotou-se essa estratégia devido a necessidade de um feedback mais rápido da situação do aluno, tendo em vista que os dados socioeconômicos estão disponíveis ainda antes do aluno iniciar as atividades letivas na instituição. Nesses testes, o algoritmo *Naive* Bayes obteve uma acurácia total de 78%, porém com taxas de

[1]Nesta abordagem o conjunto de dados original é dividido em k subgrupos disjuntos (denominados *folds*), em que para k-1 subgrupos é feito o treinamento do modelo, que é testado com o subconjunto restante. O processo é repetido usando-se subconjuntos diferentes para treinamento e teste, até completar k repetições. Ao término dos testes, calcula-se a média das taxas de acerto e erro de todos os testes.

Tabela 2. Atributos socioeconômicos levados em consideração pela classificação

Atributo	Descrição	Tipo
sexo	Sexo do estudante	Nominal (Masculino ou Feminino)
cor_pele	Cor da pele do estudante	Nominal (Branca; Negra; Indígena; Amarela; Parda)
estado_civil	Estado civil do estudante	Nominal (Solteiro; Casado; Viúvo; Separado; Outro)
renda_familiar	Renda familiar do estudante	Nominal (de 1 a 2 salários mínimos; de 10 a 20 salários mínimos; de 2 a 3 salários mínimos; de 20 ou mais salários mínimos; de 3 a 5 salários mínimos; de 5 a 10 salários mínimos; menos de 1 salário mínimo)
trabalha	Situação de trabalho (vínculo empregatício) do estudante	Nominal (não; sim, às vezes; sim, em tempo integral (mais de 30 horas semanais); sim, em tempo parcial (até 30 horas semanais))
possui_pc	Se o estudante possui computador próprio ou não	Nominal (Sim ou Não)
acessa_internet	Se o estudante tem acesso à Internet	Nominal (Sim ou Não)
sit_pai	Situação empregatícia do pai do estudante	Nominal (é aposentado; está desempregado; está trabalhando; outra; vive de rendimentos financeiros (aluguéis, aplicações bancárias, etc.;)
sit_mae	Situação empregatícia da mãe do estudante	Nominal (é aposentado; está desempregado; está trabalhando; outra; vive de rendimentos financeiros (aluguéis, aplicações bancárias, etc.)
prof_pai	Tipo de profissão exercida pelo pai do estudante	Nominal (Alto cargo político ou administrativo e assemelhados; Diretor ou Gerente, Proprietário de empresa de porte médio e assemelhados; Ocupação desconhecida; Ocupações do lar e assemelhadas; Profissional Liberal e demais profissões de nível superior; Profissões manuais não-especializadas e assemelhadas; Profissões não-manuais de rotina, Supervisor de trabalho manual, Profissões manuais especializadas e assemelhados; Proprietário de grande empresa e assemelhados; Supervisor ou inspetor de ocupações não-manuais, proprietário de pequena empresa e assemelhados)
prof_mae	Tipo de profissão exercida pela mãe do estudante	Nominal (Alto cargo político ou administrativo e assemelhados; Diretor ou Gerente, Proprietário de empresa de porte médio e assemelhados; Ocupação desconhecida; Ocupações do lar e assemelhadas; Profissional Liberal e demais profissões de nível superior; Profissões manuais não-especializadas e assemelhadas; Profissões não-manuais de rotina, Supervisor de trabalho manual, Profissões manuais especializadas e assemelhados; Proprietário de grande empresa e assemelhados; Supervisor ou inspetor de ocupações não-manuais, proprietário de pequena empresa e assemelhados)
instr_pai	Grau de instrução educacional do pai do estudante	Nominal (Curso universitário completo; Curso universitário incompleto; Ensino Fundamental (antigo 1º grau) completo; Ensino Fundamental (antigo 1º grau) incompleto; Ensino Médio (2º grau) completo ou equivalente; Ensino Médio (2º grau) incompleto; não frequentou escola; Pós-Graduação (mestrado e/ou doutorado))
instr_mae	Grau de instrução educacional da mãe do estudante	Nominal (Curso universitário completo; Curso universitário incompleto; Ensino Fundamental (antigo 1º grau) completo; Ensino Fundamental (antigo 1º grau) incompleto; Ensino Médio (2º grau) completo ou equivalente; Ensino Médio (2º grau) incompleto; não frequentou escola; Pós-Graduação (mestrado e/ou doutorado))

VN (alunos concluintes classificados corretamente) e FP (alunos concluintes classificados como não concluintes) de 10% e 90%, com valor da métrica UAC de 0,541. Valores estes muito aquém do esperado para a proposta de metodologia apresentada. Dessa forma, optou-se por utilizar as informações de desempenho dos estudantes, juntamente com os dados socioeconômicos, a fim de potencializar a eficiência da abordagem, embora a sua aplicação só possa ser feita após o término do primeiro período letivo, quando o aluno já possui as notas das disciplinas iniciais do curso.

A análise do ganho de informação dos atributos do conjunto mostra quanto cada um deles auxilia, individualmente, a classificação final dos estudantes, conforme Tabela 3. Os atributos que obtiveram uma maior influência foram as notas obtidas pelo aluno nas disciplinas do primeiro período do curso. Outra constatação interessante é que as disciplinas de exatas (“Cálculo Vetorial”, “Cálculo Diferencial e Integral I” e “Física Aplicada a Computação I”), que envolvem cálculos matemáticos, influenciam mais na decisão do classificador do que aquelas específicas da área (“Introdução a Programação” e “Introdução ao Computador”). O primeiro conjunto de disciplinas, historicamente, apresenta elevadas taxas de evasão nos cursos de Computação [de Brito et al. 2014], o que pode influenciar de forma decisiva a evasão dos alunos, principalmente naquilo que se refere ao fracasso nas disciplinas iniciais do curso [Barroso e Falcão 2004].

Tabela 3. Ranking dos atributos de acordo com seus valores de ganho de informação

Atributo	Valor do ganho de informação
nota_calculo_vet	0,2853
nota_calculo_1	0,2617
nota_fisica_1	0,2055
nota_intro_comp	0,1878
nota_intro_prog	0,1751
instr_mae	0,0368
sit_mae	0,0356
instr_pai	0,0338
prof_mae	0,0292
cor_pele	0,0271
estado_civil	0,0175

Atributo	Valor do ganho de informação
prof_pai	0,0162
turno_ens_medio	0,0156
trabalha	0,0124
escola_ens_fund	0,0101
renda_familia	0,0099
escola_ens_medio	0,0096
sit_pai	0,0061
sexo	0,0022
possui_pc	0,0010
acessa_internet	0,0007
mediageral	0

Percebe-se que o dado social mais importante é o grau de instrução da mãe do aluno (na sétima posição geral do valor de ganho de informação dos atributos), seguido por situação empregatícia da mãe, grau de instrução do pai e tipo de profissão da mãe. Esses resultados podem levar à conclusão de que dentre os dados socioeconômicos de um estudante, aqueles que se referem ao grau de instrução e profissão de seus pais são os mais importantes, sugerindo existir uma relação de continuidade de interesse acadêmico dentro da família. É possível que os pais que possuem nível de escolaridade superior completo tendam a refletir esse interesse para os seus filhos, que acabam sendo estimulados a concluírem um curso superior. Apesar de terem influência menor, não se pode desconsiderar os atributos de cunho social, pois em muitos países, principalmente no Brasil, existe uma grande desigualdade social, o que acarreta numa diferença de oportunidades e de acesso ao ensino e educação, criando uma disparidade entre alunos dentro de uma mesma instituição acadêmica.

Nesta análise destaca-se o valor (0) obtido pelo ganho de informação do atributo *mediageral*, que indica que o mesmo não possui importância para a definição da situação final do aluno no curso, impossibilitando a definição de uma fronteira clara entre as classes. Esse fato pode estar relacionado à possibilidade que alunos com bom desempenho na prova de ingresso não tenham identificação com o curso e optem por abandoná-lo, além disso é possível que alunos com desempenho inferior no ingresso, superem as suas limitações e consigam concluir o curso. Uma análise detalhada das notas obtidas na prova de ingresso, considerando isoladamente a média de cada matéria, pode ser interessante para investigar com mais detalhes esse fato. Por não ter importância, o atributo foi eliminado para os testes de classificação.

A segunda análise da estratégia avalia a capacidade de identificação dos estudantes evadidos e concluintes, a partir do algoritmo *Naive* Bayes. A acurácia total do modelo produzido foi de 85,48%, o que atesta a qualidade geral das predições. As taxas de VP e VN foram 85,60% e 85,00%, respectivamente. Os resultados mostram que os estudantes potencialmente evadidos, bem como os estudantes que provavelmente concluirão o curso, são identificados com elevadas taxas de acerto. Os FN e FP foram, respectivamente, de 14,40% e 15,00%, demostrando uma baixa taxa de erro na identificação dos grupos de estudantes. Por fim, quanto à métrica AUC, obteve-se o valor de 0,919, sendo o mínimo considerado 0,5 e o máximo 1, resultado que corrobora a boa qualidade da abordagem quanto a sua capacidade de diagnóstico de evasão de estudantes do curso de Computação.

As principais contribuições encontradas na estratégia de predição apresentada neste trabalho são: (1) facilitar a detecção de alunos com risco de evasão no curso de forma mais prévia possível; (2) auxiliar a criação e aplicação de novas ações e políticas educativas a fim de evitar a evasão de alunos propensos a evadir do curso; e (3) prover maior confiança ou estímulo para estudantes de acordo com sua classificação dada pelo algoritmo, levando a um maior comprometimento do mesmo, caso necessário.

5. Considerações Finais e Trabalhos Futuros

Este trabalho apresentou uma abordagem para o diagnóstico de evasão de estudantes, a partir de seus dados socioeconômicos e acadêmicos. Como forma de validação do método, um conjunto de 241 registros de estudantes concluintes e evadidos, do curso de Ciência da Computação da UFPB, foi analisado na ferramenta Weka. A relevância de cada atributo para a classificação e a capacidade preditiva do método foram avaliados. Constatou-se que as informações acadêmicas dos estudantes têm maior impacto na identificação na sua situação final do que os dados socioeconômicos. Em relação à predição da evasão de estudantes, o método *Naive* Bayes apresenta uma acurácia geral de 85,48%, indicando que a metodologia proposta é viável e sensata. Espera-se que o procedimento aqui apresentado possa ser adaptado à realidade de cada instituição de ensino e utilizado por alunos, professores e coordenadores de cursos para um melhor entendimento da evasão e para a criação de políticas de combate ao fenômeno no ensino superior, de maneira que os altos índices do problema sejam combatidos.

Como trabalho futuro, acredita-se que o uso do método pode ser ampliado para cursos de outras áreas do conhecimento (Humanas, Saúde, etc.), de forma a identificar novas relações, padrões, características e conhecimentos inerentes a cada área. No entanto, é preciso realizar uma análise posterior a fim de conferir a manutenção dos resultados obtidos neste trabalho. Outra abordagem seria a aplicação e a inclusão de dados demográficos dos estudantes, com objetivo de encontrar relação entre a distância da sua residência e a universidade e o meio de transporte utilizado pelo mesmo e se essas e outras informações semelhantes influenciariam o método na classificação dos estudantes.

Referências

Ang, S. L., Ong, H. C., e Low, H. C. (2016). Classification Using the General Bayesian Network. *Pertanika Journal of Science & Technology*, 24(1).

Aparecida, C., Baggi, S., e Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica.

Baggi, C. A. d. S. e Lopes, D. A. (2010). Evasão e avaliação institucional no ensino superior: Uma discussão bibliográfica. 2010.

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7:112–118.

Baradwaj, B. K. e Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

Barroso, M. F. e Falcão, E. B. (2004). Evasão Universitária: O caso do Instituto de Física da UFRJ. *Encontro Nacional de Pesquisa em Ensino de Física*, 9:1–14.

Bhardwaj, B. K. e Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

de Brito, D. M., de Almeida Júnior, I. A., Queiroga, E. V., e do Rêgo, T. G. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 25, page 882.

de Oliveira Júnior, J. G., Noronha, R. V., e Kaestner, C. A. A. (2014). Análise da correlação da evasão de cursos de graduação com o empréstimo de livros em biblioteca. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3, page 601.

de Souza, S. L. (2008). *Evasão no Ensino Superior: Um Estudo utilizando a Mineração de Dados como ferramenta de Gestão do Conhecimento em um Banco de Dados referente à Graduação de Engenharia.* Dissertação de Mestrado, Universidade Federal do Rio de Janeiro.

Dekker, G. W., Pechenizkiy, M., e Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*.

Faceli, K., Lorena, C. A., Gama, J., e Carvalho, André, C. P. L. F. (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

Gisi, M. L. (2006). A Educação Superior no Brasil e o caráter de desigualdade do acesso e da permanência. *Diálogo Educacional, Curitiba*, 6(17):97–112.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., e Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Osmanbegović, E. e Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review*, 10(1).

Pascoal, T. A., de Brito, D. M., e do Rêgo, T. G. (2015). Uma abordagem para a previsão de desempenho de alunos de Computação em disciplinas de programação. In *Nuevas Ideas en Informática Educativa TISE 2015*, pages 454–458.

Prietch, S. S. e Pazeto, T. A. (2010). Estudo sobre a Evasão em um Curso de Licenciatura em Informática e Considerações para Melhorias. *WEIBASE, Maceió/AL*.

Romero, C. e Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146.

Romero, C. e Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618.

Silva, L. A., Morino, A. H., e Sato, T. M. C. (2014). Prática de mineração de dados no exame nacional do ensino médio. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3, page 651.

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., e Lobo, M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132):641–659.

Vitelli, R. F., Rocha, C. S., e Fritsch, R. (2011). Estudo sobre evasão nos cursos de graduação de uma Instituição de Ensino Superior privada: aplicação de regressão logística.

Witten, I. H. e Frank, E. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xenos, M. (2004). Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education*, 43(4):345–359.