Rede Bayesiana para Previsão de Evasão Escolar

William Maria[1], João L. Damiani[1], Max Pereira[1]

[1]Núcleo de Inteligência Artificial e Análise de Dados
Universidade do Sul de Santa Catarina (UNISUL) – Tubarão, SC – Brasil

willsilvano@gmail.com, jlucasd01@gmail.com, max.pereira@unisul.br

Abstract*. Given the problem of school evasion, a problem that impacts educational institutions around the world, this paper presents the application of Bayesian networks to predict the percentage chance of student evasion, in order to assist educational managers in preventing these types of situations. The prediction is performed based on the characteristics of the students, collected from the data base system used by SENAI (Tubarão/SC). It is possible to manipulate these characteristics, by the manager, in order to simulate scenarios to minimize the chances of the student evasion. Through the validation of the results, we have obtained 85.6% of accuracy, which indicates a good performance of the Bayesian network modeled for the system developed.*

Resumo*. Tendo em vista o problema de evasão escolar, problema este que impacta instituições de ensino do mundo inteiro, este artigo apresenta a aplicação de Redes Bayesianas, com o intuito de predizer os percentuais de chance de evasão dos alunos, com o objetivo de auxiliar os gestores educacionais na prevenção destes tipos de situações. A predição é realizada com base nas características dos alunos, coletadas do sistema utilizado pelo SENAI de Tubarão/SC. É possível ainda manipular tais características, por parte do gestor, a fim de simular cenários com o objetivo de minimizar as chances de o aluno evadir. Através da validação dos resultados foi obtido 85,6% de taxa de acerto, o que indica um bom desempenho da rede bayesiana modelada para o sistema desenvolvido.*

1. Introdução

A evasão escolar é um problema recorrente que atinge as instituições de ensino e que possui proporções de níveis mundiais. Tal tema vem sendo foco de diversos estudos e pesquisas no meio educacional, segundo Hipolito et al. (2007). Diversas situações podem ser caracterizadas como evasão, tais como: o trancamento de um curso por um estudante, a desistência por falta de interesse, a falta de recursos financeiros do aluno, motivo de doença, gravidez precoce, ou até mesmo a desistência devido à incompatibilidade de horários das aulas com o mercado de trabalho ou ainda quando os estudantes dão inicio á carreira profissional.

Os números da evasão escolar no país têm sido um fator preocupante para as instituições de ensino, gestores e governantes, chegando a um percentual de 24,3% no ano de 2012 (Paim, 2013). O governo entre os anos de 2004 e 2011 minimizou a taxa de evasão no país e tem olhado de uma forma mais cautelosa para este tema, visto que não é somente o meio educacional que sai prejudicado com o alto índice de evasão, mas também todo o meio socioeconômico. Muito se tem estudado a respeito desse problema como, por

DOI: 10.5753/cbie.wcbie.2016.920

exemplo, os trabalhos realizados por Kotsiantis et al. (2003), além de Hämäläinen et al. (2004).

Tendo em vista o quão importante é a permanência de um estudante em uma instituição de ensino, foi desenvolvido um sistema computacional utilizando redes bayesianas, capaz de predizer o percentual potencial de evasão dos alunos, permitindo ainda que o gestor possa simular os possíveis cenários para o aluno, a fim de minimizar as chances de evasão escolar na instituição. Nesse sentido, as redes bayesianas possuem papel fundamental na obtenção da probabilidade potencial de evasão do aluno. Assim, segundo Russel (1995) através das probabilidades, pode-se elucidar com níveis de certeza problemas com base em evidências de uma situação.

Para estudo de caso foi utilizada a instituição Serviço Nacional de Aprendizagem Industrial (SENAI) na unidade de Tubarão / SC, onde também foi realizada a coleta dos dados necessários para a construção da rede bayesiana, bem como para a alimentação de dados para o sistema.

2. Trabalhos correlatos

Nassar et al. (2004), apresentaram o desenvolvimento de um sistema de gestão do fenômeno da evasão discente utilizando a modelagem de redes bayesianas. A representação é feita em um grafo direcionado acíclico, cujos nós de entrada representam os fatores que interferem na evasão e o nó de saída os possíveis resultados de um aluno em determinado curso. Foram realizadas simulações com base no teorema de bayes que permitem estimar o risco de evasão em determinado curso, a partir do conhecimento histórico de evasão e de fatores pessoais do discente.

Hämäläinen et al. (2004) analisaram duas disciplinas de programação de computadores em um curso online. O trabalho utilizou regras de associação e modelos probabilísticos para identificar os fatores mais importantes para predizer os resultados finais nas duas disciplinas. Kotsiantis et al. (2003) compararam diversos algoritmos para detectar o mais adequado para predizer a evasão dos alunos.

Pode-se citar ainda o trabalho realizado por Dekker et al. (2009), onde os autores analisaram dados de alunos de graduação do curso presencial de Engenharia Elétrica da universidade de Eindhoven. Neste trabalho, identificou-se já no primeiro ano letivo os alunos com risco de evasão. Os autores avaliaram diversos algoritmos da ferramenta de mineração de dados Weka (Hall et al. 2009) afim de detectar o mais adequado. O experimento analisou diversos dados dos alunos e obteve um percentual de 75% a 80% de precisão com o classificador de árvore de decisão.

3. Materiais e métodos

3.1 Ferramentas

Para realizar o desenvolvimento do software foi utilizada como linguagem de programação Java, com auxilio da IDE Eclipse, bem como o framework JSF 2.0, além de utilizar a suíte de componentes Primefaces 5.0. Para o banco de dados foi utilizado o banco de dados relacional PostgreSQL 9.3 e o framework de mapeamento e persistência Hibernate 4.3.8.

Para modelagem e testes da rede bayesiana foi utilizada a ferramenta Genie, uma ferramenta gratuita e livre de limitações. Foi utilizado, também no desenvolvimento, a API JSmile[1], responsável por realizar a inferência da rede modelada com o Genie.

Os dados de teste utilizados na aplicação foram extraídos da base de dados do SGN (Sistema de Gestão de Negócios) utilizado pelo SENAI/SC. Foram extraídos apenas dados dos cursos técnicos da unidade do SENAI de Tubarão/SC. A importação destes dados é feita através do consumo de um webservice, que faz a leitura dos dados necessários no banco de dados do sistema SGN e os disponibilizam em formato de objetos JSON.

3.2 Método computacional

3.2.1 Modelagem da rede bayesiana

Para a construção do modelo da rede bayesiana proposta como solução, foi necessário efetuar um levantamento de quais seriam as principais informações que poderiam ser relacionadas aos alunos para posterior predição da evasão escolar.

Inicialmente foi analisada a base de dados do Sistema de Gestão de Negócios (SGN), que é utilizado pela unidade do SENAI de Tubarão, a fim de encontrar informações importantes que pudessem ser adicionadas como nós no modelo da rede bayesiana. Depois de uma análise dessas informações, foi elencado as possíveis variáveis candidatas para compor o modelo da rede bayesiana.

Após o levantamento das variáveis, foi elaborado um modelo prévio da rede e apresentado para o coordenador do curso técnico e para a coordenadora pedagógica da unidade do SENAI estudada. O resultado final da modelagem da rede, o qual foi validado com os coordenadores consta na figura 1.

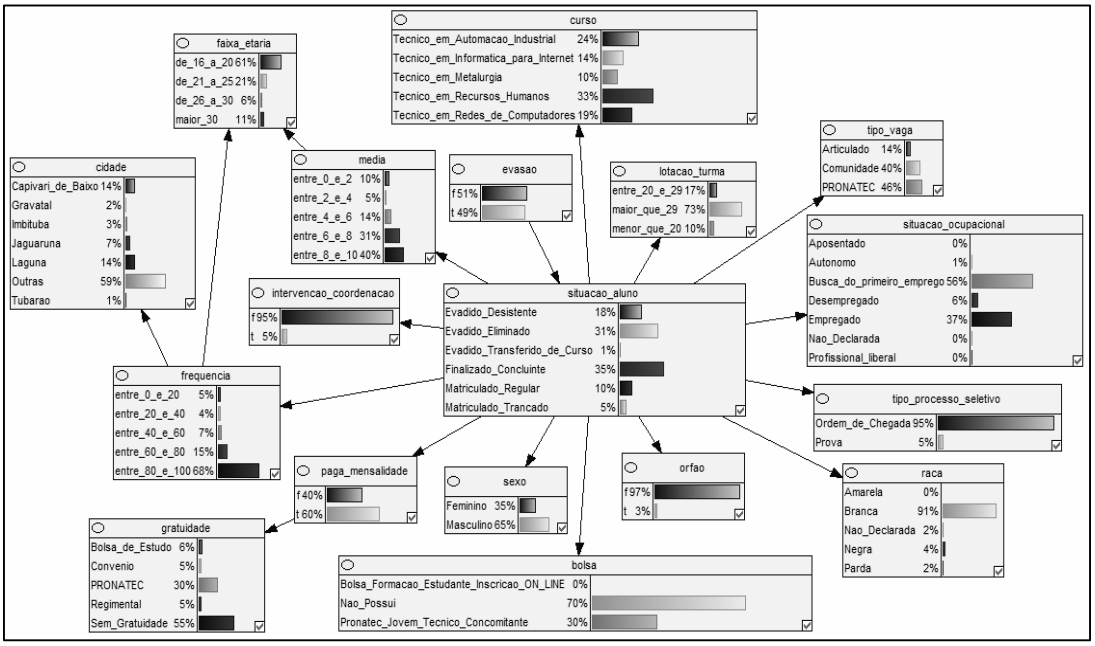

Figura 1. Modelo da rede bayesiana.

[1] Download e documentação disponível em https://dslpitt.org/genie/wiki/SMILE_Documentation

As duas principais informações no modelo da rede são os nós "evasao" e "situacao_aluno", pois estas são as saídas da rede, ou seja, informações que o sistema disponibilizará, indicando ao gestor as chances de evasão dos alunos. O primeiro nó caracteriza os percentuais de evasão dos alunos. Já o segundo nó caracteriza o percentual para cada situação em que o aluno poderá estar.

A figura 2 representa um exemplo do nó "evasao" na rede bayesiana. Pode-se observar que, de acordo com as informações dessa figura, o aluno teria 52% de probabilidade de não evadir e 48% de probabilidade de evadir do curso.

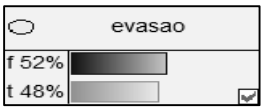

Figura 2. Nó que representa a evasão dos alunos.

A figura 3 representa um exemplo do nó "situacao_aluno" na rede bayesiana. Pode-se observar que, de acordo com as informações dessa figura, o aluno teria 31% de probabilidade de evadir do curso com situação "Evadido_Eliminado", 18% de evadir com situação "Evadido_Desistente", 1% de evadir na situação "Evadido_Transferido_de_Curso", 10% de se manter regular no curso, 5% de efetuar um trancamento e 35% de concluir o curso.

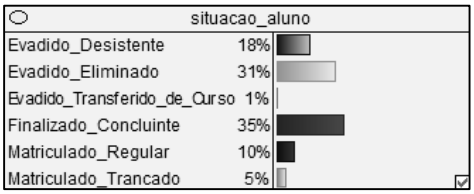

Figura 3. Nó que representa as situações dos alunos.

As dependências condicionais entre os nós da rede podem ser verificados na figura 1, sendo estas representadas pelas setas que ligam os nós entre si. O nó de onde sai a seta é dependente do nó que recebe a seta. Desta forma pode-se verificar que o nó "evasao" é dependente do nó "situacao_aluno", e que este, por sua vez, é dependente de vários outros nós, como por exemplo os nós: "sexo", "media", "frequencia", dentre outros.

A tabela 1 representa a tabela de probabilidade incondicional do nó "evasao" para a rede bayesiana modelada e com base no número de alunos importados para o sistema. O total de alunos importados foi de 666. Deste universo, 337 evadiram dos seus cursos e 329 não evadiram. Colocando os valores nos cálculos para posterior preenchimento da tabela de probabilidade do nó "evasão", ficaria 337/666 para evasão = t (Verdadeiro) e 329/666 para evasão = f (Falso).

▶	f	0.514
	t	0.486

Tabela 1. Tabela de probabilidade incondicional do evento evasão.

A tabela 2 representa a tabela de probabilidades do nó "situacao_aluno", da rede bayesiana modelada, condicionada pelo nó "evasao", desta forma pode-se dizer que o nó "situacao_aluno" é dependente do nó "evasao", e a notação que representa essa dependência é *P(situacao_aluno|evasao)*.

	evasao	f	t
▶	Evadido_Desistente	0	0.36625514
	Evadido_Eliminado	0	0.63374486
	Evadido_Transferido_de_Curso	0.01945525	0
	Finalizado_Concluinte	0.6848249	0
	Matriculado_Regular	0.19455253	0
	Matriculado_Trancado	0.10116732	0

Tabela 2. Tabela de probabilidade condicional do nó "situacao_aluno".

No sistema, o teorema de bayes foi utilizado para efetuar o cálculo de probabilidades condicionadas onde existiam uma ou mais ligações entre os nós de evidências e hipótese.

Como exemplo, no modelo de estudo, podemos citar os nós "frequencia" e "media", ambos dependentes do nó "faixa_etaria". Nesse caso, para realizar os cálculos e preencher a tabela de probabilidades do nó "faixa_etaria", foi aplicado o teorema de bayes, conforme representado na equação 1. Na figura os "*n*" representam cada um dos estados que compõem os nós "faixa_etaria" (*FE*), "frequência" (*FR*) e "media" (*MD*).

$$P(FEn|FRn, MDn) = \frac{P(FEn).P(FRn|FEn).P(MDn|FEn)}{P(FRn, MDn)} \quad \text{Eq. 1}$$

Após a conclusão da modelagem da rede bayesiana, tendo já definidos os nós e as dependências entre os mesmos, a etapa seguinte visa calcular e definir as tabelas das probabilidades condicionais e incondicionais da rede. É com base nessas tabelas que as inferências são realizadas sobre a rede, retornando os percentuais possíveis de evasão dos alunos nos cursos. Para realizar as inferências foi utilizado a API Jsmile, que por padrão se utiliza do algoritmo de Lauritzen-Spiegelhalter (Cowell et al, 1999).

4. Ferramenta para predição de evasão

O sistema SPEED (Sistema Preditivo de Evasão Escolar Discente) serve de auxilio aos gestores na exposição da situação potencial de evasão de alunos nos cursos aos quais estão vinculados e desta forma servindo de monitoramento para posteriores medidas pedagógicas a serem tomadas por estes gestores. As principais funcionalidades do sistema estão dispostas nos tópicos seguintes.

4.1 Importação de alunos

Na tela de importação de alunos o gestor irá efetuar a chamada do webservice o qual irá importar os dados dos alunos da fonte de informações.

4.2 Listagem de alunos

Na tela de listagem de alunos o gestor poderá selecionar dentre as turmas e cursos para filtrar os alunos a serem analisados.

Para cada aluno listado, o sistema carregará as evidências (variáveis) deste para a rede bayesiana, de forma que cada evidência do aluno seja definida como verdadeira, ou seja, com percentual 100%. O sistema envia estas informações para a API Jsmile, e esta por sua vez efetua os cálculos para obtenção das probabilidades de cada nó da rede

bayesiana modelada. Como o foco é saber o percentual de evasão dos alunos, o sistema exibe o percentual para o estado "t" (sim) do nó "evasão", ou seja, o percentual estimado do aluno evadir da turma/curso selecionado.

4.3 Simulação de situações de evasão

O gestor seleciona um aluno da lista, a fim de verificar as características deste aluno e efetuar simulações no cenário de evasão para o aluno selecionado. Tais simulações tem o objetivo de verificar quais características, para um determinado aluno, possuem maior impacto na sua possível evasão ou não. Ao acessar essa tela, representada na figura 4, o sistema carrega todas as variáveis deste aluno que compõem o modelo da rede. Algumas das informações são fixadas, sem possibilidade de simulação, como por exemplo: o sexo, a idade, raça, etc. Outras informações ficam habilitadas para que o gestor possa efetuar simulações, como por exemplo, a média, a frequência, tipo de vaga, dentre outras informações. Ao selecionar as informações que irão compor o cenário de simulação, o sistema define como verdade cada uma das informações que o gestor selecionou e retorna os percentuais para os nós “evasao” e “situacao_aluno”. Desta forma o gestor poderá comparar as características que o aluno possui atualmente com as características que foram simuladas a fim de verificar se com a simulação houve melhora e assim agir nos pontos necessários.

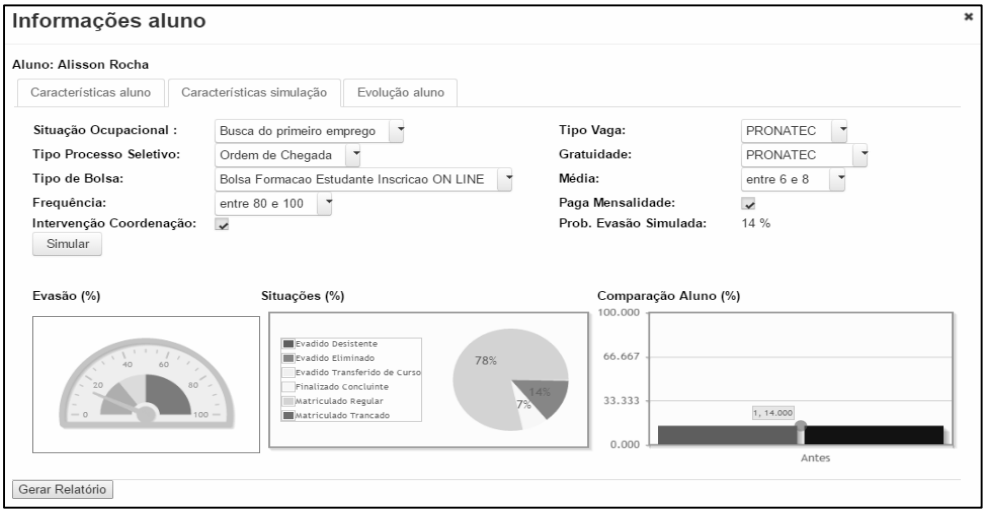

Figura 4. Tela de simulação de evasão dos alunos no sistema SPEED.

4.4 Evolução do percentual de evasão dos alunos

Nesta tela do sistema o gestor poderá acompanhar a evolução do percentual de evasão inferido pela rede bayesiana, ao longo das importações realizadas no sistema.

A cada importação realizada no sistema é calculado o percentual de evasão de cada um dos alunos importados e salvo este valor na base de dados, mantendo um histórico para cada aluno.

Essa informação é importante para que o gestor possa acompanhar se as medidas que estão sendo tomadas para evitar a evasão do aluno estão surtindo efeito ou não, e com base nessa evolução, tomar as medidas cabíveis para a situação.

5. Resultados e discussões

Para validação da eficiência da rede bayesiana implementada, foram feitos alguns testes. Há dois tipos de cálculos realizados por uma rede bayesiana: a atualização de crenças e a revisão de crenças. A atualização é o cálculo de probabilidades das variáveis aleatórias e a revisão refere-se à obtenção das probabilidades das hipóteses diagnósticas (nó de saída) e a identificação da hipótese diagnóstica com maior valor de probabilidade.

Para o experimento foi utilizada a técnica de validação cruzada com o método k-fold, onde foram selecionados, aleatoriamente, 100 alunos dentre o total de 666 alunos contidos na base de dados. Para cada um dos 100 alunos foram setadas as evidências da rede, de acordo com as variáveis do aluno (frequencia, media, faixa_etaria, sexo, curso, etc), com exceção do nó "evasao", que é o resultado que se almeja alcançar para comparação.

Após atribuir as evidências na rede bayesiana, foi realizado o processo de inferência da rede para o nó "evasao", verificando o percentual apresentado para o estado "t" (verdadeiro) e para o estado "f" (falso). Logo após, foi selecionado o estado com o maior percentual para este nó. Por exemplo, se a inferência da rede para o aluno retornou 40% para o estado "t" e 60% para o estado "f", então entendemos que o resultado da rede para este aluno foi a não evasão (60%).

Este resultado de evasão / não evasão apresentado pela rede, foi comparado com a informação contida na base de dados, proveniente da importação de alunos, relacionando se o aluno evadiu ou não no sistema SGN.

Esse processo foi repetido 10 vezes, selecionando 100 alunos aleatoriamente em cada iteração, e ao final foi calculada a média de acerto e erro com base nas médias obtidas nas 10 iterações. As médias obtidas foram, respectivamente, 85,6% para acerto e 14,4% para erro.

Para a avaliação dos resultados obtidos com o experimento, foi utilizada uma matriz de confusão, conforme consta na tabela 3. Nessa matriz constam as quantidades de acertos e erros com relação às inferências realizadas pela rede. A matriz de confusão mostra o número de classificações corretas em oposição às classificações preditas para cada classe, ou neste caso, para cada estado do nó "evasao". Na matriz apresentada na tabela 3 as linhas representam as classes possíveis para o nó "evasao" (Sim e Não) e as colunas representam as classificações inferidas pela rede bayesiana.

Na matriz de confusão, a diagonal principal representa os acertos da rede, sendo 379 acertos para evasão e 477 acertos para não evasão. A diagonal inversa representa os erros preditos pela rede, sendo que em 123 casos a rede inferiu não evasão para alunos que evadiram (falsos negativos), e em 21 casos a rede inferiu evasão para alunos que não evadiram (falsos positivos).

qtd.	SIM	NÃO
SIM	379	123
NÃO	21	477

Tabela 3. Matriz de confusão das 10 iterações com 100 alunos aleatórios.

As taxas de precisão são as taxas de acerto da rede sobre os casos que realmente aconteceram. Para os casos de evasão a taxa de precisão foi de 75,5%, e para os casos de não evasão a precisão foi de 95,78%.

O cálculo das taxas de precisão está representado nas equações 2 e 3, onde *Tp* e *Tn* representam, respectivamente, os acertos inferidos para evasão e não evasão; *Fn* e *Fp* representam, respectivamente, os erros inferidos para evasão e não evasão.

$$\frac{T_p}{T_p + F_n} \quad \text{Eq. 2}$$

$$\frac{T_n}{F_p + T_n} \quad \text{Eq. 3}$$

O resultado final do desempenho da rede bayesiana modelada ao final do experimento obteve uma taxa de 85,6% de acerto e 14,4% de erro.

Para os casos em que a rede errou a inferência, em comparação com a real saída, foi realizada uma análise detalhada, averiguando os motivos para este comportamento na rede bayesiana modelada. Os casos avaliados foram coletados durante o experimento anteriormente citado.

Nesta avaliação, os "erros" foram classificados de duas formas: erros "sim" e erros "não". Os chamados erros "sim", foram os casos onde o aluno evadiu de fato e a rede acusou como não (atribuindo um baixo percentual de evasão ao aluno). Os chamados erros "não", foram os casos onde os alunos não evadiram e a rede acusou que sim, atribuindo-lhes um alto índice para evasão.

Nesta análise foram verificados de forma minuciosa os "acompanhamentos pedagógicos" (registros acadêmicos feitos pela coordenação do curso) dos alunos selecionados nesta amostra aleatória, além de uma análise do desempenho até o momento da evasão (nos casos em que foram concretizadas). Essas informações foram analisadas para precisar o real motivo da evasão e avaliar se foi um erro crítico cometido pela rede bayesiana ou se foram fatos atípicos não previstos pela rede.

Para os casos de erro "não" foram analisados 9 casos de alunos nesse cenário, e foi visto que a rede apesar do "erro", inferiu corretamente o alto percentual de evasão para os aluno que não evadiram. Em 55,56% dos casos (5 alunos) os alunos trancaram o curso, ou seja, o alto percentual indicado se refletiu no trancamento destes. Nos demais casos, 44,44% (4 alunos), a rede atribuiu um alto percentual de evasão devido ao baixo desempenho dos alunos, o que não é incorreto.

Para os casos de erro "sim" foram analisados 83 casos de alunos nesse cenário, e foi visto que tiveram situações atípicas as quais eram imprevisíveis para a rede na qual não havia possibilidades de previsão precisa para tais alunos.

Desta forma, em 22,89% dos casos (19 alunos), a rede calculou um baixo percentual para evasão, pois baseado no bom desempenho dos alunos, estes tinham uma perspectiva próspera de continuidade no curso, porém por fim acabaram evadindo em contradição as suas boas características, sendo assim evasões sem motivos claros.

Em 38,55% dos casos (32 alunos), a rede "errou" ao indicar um baixo percentual para estes alunos que vieram a evadir, sendo que eles possuíam características potenciais de evasão.

Em 30,12% dos casos (25 alunos), os alunos seguiram no curso até o período final (onde se opta por estágio ou TCC para conclusão do curso) tendo um bom desempenho até este período aonde vieram a evadir. Pode-se subentender que a rede agiu

de forma adequada, pois o bom desempenho realizado até o ultimo período reduziu o percentual geral de chances de tais alunos evadirem, o que de fato tem consistência.

Completando a análise tivemos ainda 6,03% (5 casos) e 2,41% (2 casos) de alunos que tiveram sua evasão por motivos atípicos, sendo respectivamente por motivo de saída para cursar no ensino superior e alunos que tiveram que abandonar o curso por motivo de doença pessoal ou familiar.

6. Conclusões

A evasão é um problema de proporções mundiais e que ainda traz muita dor de cabeça para as instituições de ensino, tendo várias consequências tanto para o meio educacional quanto para o social. Procurando tentar resolver esse problema, foi desenvolvido um software capaz de predizer o percentual de chance de um aluno evadir do curso.

O sistema desenvolvido alcançou uma considerável taxa de acerto nas predições de evasão com base na rede bayesiana modelada, o que é algo bom para um primeiro protótipo. Para que se alcance percentuais maiores se faz necessário realizar uma nova análise dos dados disponíveis no sistema de onde os mesmos foram coletados, e realizar ajustes na modelagem, adicionando e ajustando os nós e estados na rede bayesiana.

Apesar de ter sido aplicado na instituição de ensino SENAI de Tubarão, apenas com os cursos técnicos, a solução foi desenvolvida de forma modular, de forma que fique simples a aplicação em outras instituições de ensino, apenas tendo que criar um modelo da rede para a instituição desejada e realizar algumas pequenas modificações no sistema.

Para trabalhos futuros pretende-se deixar o sistema ainda mais dinâmico, permitindo que o próprio administrador do sistema possa criar novos nós, estados e relacionar as dependências entres os nós da rede. Também deixar a funcionalidade de importação de dados dos estudantes ainda mais dinâmica, validando os dados provenientes dos sistemas externos, de acordo com os nós e estados componentes configurados no sistema.

7. Referências

Cowell, R. G.; Dawid, A. P.; Lauritzen, S. L.; Spiegelhalter, D. J. (1999) Probabilistic Networks and Expert Systems. Springer-Verlag New York Inc., New York, NY.

Dekker G., Pechenizkiy M. and Vleeshouwers J. (2009) "Predicting Students Drop Out: A Case Study". In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, T. BARNES, M. DESMARAIS, C. ROMERO and S. VENTURA Eds., Pages 41-50.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten. I.H. (2009) "The WEKA Data Mining Software: An Update" SIGKDD Explorations, Volume 11, Issue 1.

Hämäläinen, W., Suhonen, J., Sutinen, E., and Toivonen, H. (2004) "Data mining in personalizing distance education courses". In world conference on open learning and distance education, Hong Kong, pp. 1–11.

Silva Filho, R.L.L., Motejunas, P.R., Hipólito, O. & Lobo, M.B.C.M. (2007). A evasão no ensino superior brasileiro. Cadernos de Pesquisa, 37(132), 641-659.

Kotsiantis, S., Pierrakeas, C. e Pintelas, P., (2003) "Preventing student dropout in distance learning using machine learning techniques." KES, eds. V. Palade, R. Howlett & L. Jain, Springer, volume 2774 of Lecture Notes in Computer Science, pp. 267–274

Nassar, Silvia M; Neto, Eugênio Rovaris; Catapan, Araci Hack; Pires, Maria Marlene de Souza. Inteligência Computacional aplicada à Gestão Universitária: Evasão Discente. 2004. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/35808/Silvia%20M%20Nassar1%20-%20inteligencia%20computacional.pdf?sequence=4>. Acesso em: 23 ago. 2014.

Paim, Paulo. Paim considera evasão escolar no Brasil preocupante. 2013. Disponível em <http://senado.jusbrasil.com.br/noticias/112004353/paim-considera-evasao-escolar-no-brasil-preocupante?ref=home>. Acesso em: 23 ago. 2014.