TECNOLOGIAS E FERRAMENTAS DE GERAÇÃO DE CONHECIMENTO PARA SUPORTE À DECISÃO NO ACESSO E PERMANÊNCIA AS UNIVERDIDADES: APLICAÇÃO NOS DADOS DO ENEM.

MICHELE ANDRÉIA BORGES
SILVIA MODESTO NASSAR
ARAN BEY TCHOLAKIAN MORALES

Resumo: Este artigo tem como objetivo verificar a eficácia dos Sistemas Especialistas Probabilísticos (SEP) e Redes Bayesianas para a geração de insumos de conhecimento no que se refere às iniciativas de gestão pública da educação no acesso e permanência às Universidades Federais Brasileiras. Para alcançar o objetivo proposto foi realizado um levantamento bibliográfico, metapesquisa do tema Redes Bayesianas em uma base de dados internacional eletrônica SCOPUS e uma aplicação de Redes Bayesianas em dados do Exame Nacional do Ensino Médio (Enem) para verificar o perfil dos estudantes de Santa Catarina. O tema, redes bayesianas, vem evoluindo desde o ano 1988, com um total de 12.134 publicações na área. A construção da rede bayesiana aplicada aos dados do Enem se mostrou bastante robusta com perspectiva de novas inferências para dar um suporte maior às decisões de especialistas na área de domínio.

Palavras-chave: *Sistemas Especialistas Probabilísticos, Redes Bayesianas, Exame Nacional do Ensino Médio.*

1. Introdução

Em nosso cotidiano é comum lidarmos com situações na qual a certeza de ocorrência de um determinado evento não é determinista. Por exemplo, quando o nosso time de futebol querido vai entrar em campo sabemos que podem ocorrer três situações distintas: ganhar, perder ou empatar; ou seja, não temos a certeza exata do que vai ocorrer até que os 90 minutos (mais os possíveis acréscimos) de jogo se encerrem. No entanto, esse não é o único caso de incerteza em nossas vidas. Usualmente, falamos expressões do tipo “entre”, “bom”, “ruim”, “baixo”, “alto” e assim por diante. A esse tipo de incerteza costumamos denominar de *incerteza por vagues*.

Sendo assim, como podemos representar conhecimento constituído de incerteza? Diversas áreas como, por exemplo, a medicina, é repleta de conhecimento incerto. Dessa

forma exigem-se do mundo científico ferramentas que suportem essas incertezas de maneira a se obter resultados mais realistas e positivos.

Sistemas Especialistas Probabilísticos (SEP) são ferramentas geradoras de conhecimento para dar suporte à decisão cujo, o domínio, é marcado por incertezas. Para codificar esse conhecimento incerto, redes bayesianas constituem um modelo gráfico que representa de forma simples as relações de causalidade das variáveis do sistema.

Com o intuito de verificar a eficácia dos Sistemas Especialistas Probabilísticos (SEP) e Redes Baysianas, assim como a geração de insumos para a área de gestão da Educação nos propusemos, neste artigo, a uma aplicação dos SEP em dados do Exame Nacional do Ensino Médio (Enem).

Este artigo esta organizado da seguinte forma: uma breve descrição e caracterização sobre incertezas. Conceitualização de Sistemas Especialistas Probabilístico. Probabilidade Bayesiana: Teorema de Bayes e Redes Bayesianas. Aplicação da teoria de Bayes em dados do Enem. Finalmente, os resultados e as conclusões.

Para alcançar o discernimento dos conteúdos deste artigo foi feito um levantamento bibliográfico e metapesquisa do tema Redes Bayesianos realizado em uma base de dados internacional eletrônica SCOPUS.

2. Raciocínio sobre incertezas

Informação real é geralmente imperfeita e incompleta ou, podemos assim dizer, tomada por incertezas na qual precisamos dar um tratamento diferenciado da lógica booleana, que admite somente valores verdadeiros (comumente representando pelo número 1) ou falsos (comumente representando pelo número 0), para que se consiga fazer a representação correta da informação a fim de se obter conhecimento que de suporte as decisões de nossas vidas.

Em conformidade com Klir e Folger (1998) “Incerteza origina-se de alguma deficiência da informação. A informação pode estar incompleta, ser vaga, imprecisa ou contraditória.”. E, para dar tratamento a essas incertezas existem teorias como *fuzzy sets*, probabilidade, teoria da evidência, entre outras. A Figura 2.1 ilustra essas colocações com a *Taxonomia da Ignorância* denominada por Bracarense e Nassar (2010).

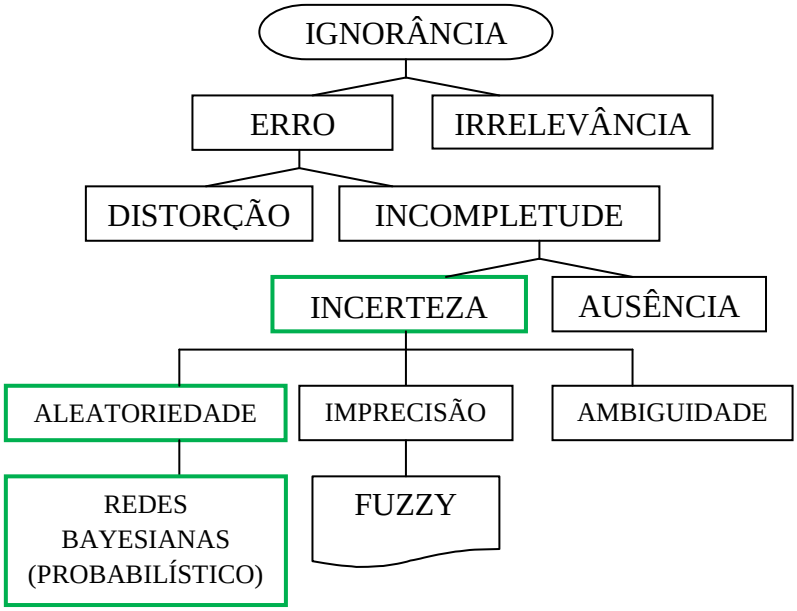

Figura 2.1: Taxonomia da Ignorância (BRACARENSE e NASSAR, 2010)

A principal vantagem do raciocínio probabilístico ou *fuzzy* sobre raciocínio lógico Booleano é o fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará. (CHARNIAK, 1991)

Neste artigo estamos interessados em investigar incertezas do tipo aleatórias utilizando a ferramenta de Redes Bayesianas na aplicação proposta.

3. Sistemas Especialistas Probabilísticos

Sistemas Especialistas (SE) tem sua raiz na área de Inteligência Artificial, nascida na década de 50. Os SE são programas de computadores que imitam o comportamento de especialistas humanos dentro de um domínio de conhecimento específico. (LIEBOWITZ, 1988)

As principais características de um Sistema Especialista, segundo Liebowitz (1988), são: a habilidade para trabalhar ao nível do especialista; representar o conhecimento específico de um determinado domínio; incorporar o processo de explanação e formas de manipulação de incerteza e normalmente pertinente a problemas que podem ser simbolicamente representados.

Um sistema especialista é composto por um banco de conhecimento (ou base de conhecimentos) e um conjunto de mecanismos de inferência (PANTALEÃO, 2003). De acordo com Ramos (1995) a base de conhecimento é talvez o mais importante componente do sistema especialista. A maioria dos sistemas é composta de regras empíricas baseadas na experiência do(s) especialista(s) consultado(s).

A base do conhecimento é composta por duas etapas: *aquisição de conhecimento* e *representação de conhecimento.* A etapa de aquisição do conhecimento constitui a fase mais complexa do desenvolvimento do sistema especialista (RAMOS, 1995). Esta por sua vez, estabelece o processo à extração de fatos e a heurística, associada com a tarefa e seu ambiente, do especialista (LIEBOWITZ,1988). A representação do conhecimento é na verdade a formalização por meio de métodos da etapa de aquisição do conhecimento e, como sugere o título deste terceiro item, o método utilizado para a representação do conhecimento é o Probabilístico.

4. Probabilidade Bayesiana

A probabilidade é o ramo da matemática que visa à formulação de modelos teóricos, abstratos, para o tratamento matemático da ocorrência (ou não ocorrência) de fenômenos aleatórios; em termos sucintos, pode caracterizar-se como a Matemática do acaso, da incerteza. (FREITAS, 2009)

No ano de 1701, nascia em Londres, Thomas Bayes. Seu fascínio e dedicação ao estudo do aleatório deram origem ao que chamamos de Teoria de Bayes. (MLODINOW, 2009)

A teoria de Bayes trata essencialmente do que ocorre com a probabilidade de ocorrência de um evento se outros ocorrerem, ou dado que ocorram. Para Mlodinow (2009), “todos nós fazemos julgamentos bayesianos”. Um exemplo singelo disso é, numa família com duas crianças, qual é a probabilidade de que, se uma delas for menina, ambas sejam meninas? O *se* atribuído nessa pergunta faz com que este seja um problema de probabilidade condicional, ou seja, do que especifica a Teoria de Bayes.

Hoje a análise bayesiana é amplamente empregada na ciência e na indústria, (MLODINOW, 2009). No entanto, algumas considerações em relação ao emprego desta teoria devem ser tomadas. Pois, a probabilidade de que um evento A ocorra se um evento B ocorrer geralmente difere da probabilidade de que um evento B ocorra se um evento A ocorrer. Não levar esse fato em consideração pode-se induzir ao erro.

4.1. Teorema de Bayes

Dado dois eventos E e F tal que $P(E) \neq 0$ e $P(F) \neq 0$, temos que a probabilidade do evento F ocorrer se o evento E ocorre, é dada pela equação:

$$P(F|E) = \frac{P(E|F) \cdot P(F)}{P(E)} \tag{4.1.1}$$

Além disso, dado n eventos mutuamente exclusivos e exaustivos $F_1, F_2, \ldots, F_n$ tais que $P(F_i) \neq 0$ para todo i, temos que para $1 \leq i \leq n$:

$$P(F_i|E) = \frac{P(E|F_i) \cdot P(F_i)}{\sum_{j=1}^{n} P(F_j) \cdot P(E|F_j)} \tag{4.1.2}$$

Outro aspecto importante de cálculos probabilísticos é a *independência de eventos*. A idéia básica subjacente ao conceito probabilístico de independência entre dois eventos é que o conhecimento de certa informação sobre um evento não traz informação adicional sobre o outro. Isto é, se e somente se, ao saber que o evento E_1 ocorreu isto não trouxer informação sobre o evento E_2, e ao saber que o evento E_2 ocorreu isto não trouxer informação sobre o evento E_1, então se diz que ocorre a independência entre estes eventos. A equação matemática que expressa essa independência de eventos é dada por:

$$P(E_1 \wedge E_2) = P(E_1 \cdot E_2) \tag{4.1.3}$$

Contudo, é possível inferir o teorema de Bayes para evidências múltiplas e independentes a partir da equação:

$$P\left(F_j \middle| E_1 \wedge E_2 \wedge \ldots \wedge E_n\right) = \frac{P(F) \cdot \prod_{i=1}^{n} P(E_i|F_j)}{P(E_1 \wedge E_2 \wedge \ldots \wedge E_i \wedge \ldots \wedge E_n)} \tag{4.1.4}$$

A equação 4.1.4 é comum nos domínios de aplicação de Sistemas Especialistas Probabilísticos (SEP) devido à existência de várias hipóteses concorrentes cada uma com um conjunto distinto de evidências.

4.2. Redes Bayesianas

Uma rede bayesiana é um grafo acíclico direcionado (DAG), ou seja, é orientado e sem ciclos conforme ilustra a Figura 4.2.1, que codifica relações probabilísticas entre as distinções de interesse em um problema de raciocínio incerto, (PEARL, 1988). Do ponto de vista de um especialista, Redes Bayesianas constituem um modelo gráfico que representa de forma simples as relações de causalidade das variáveis de um sistema. (MARQUES e DUTRA, 2008)

A representação formal do conhecimento, isto é, a topologia da rede é composta por partes *qualitativas* e *quantitativas.* Nassar (1998) define como sendo:

- *Parte Qualitativa:* Representa o modelo gráfico (grafo acíclico direcionado), ou seja, as variáveis, também chamadas de *nós*, e as regras que são as relações de dependência condicional entre essas variáveis representadas pelos os arcos direcionados;
- *Parte Quantitativa:* É o conjunto de probabilidades condicionais associadas aos arcos existentes no modelo gráfico e as probabilidades estimadas à priori das hipóteses diagnósticas.

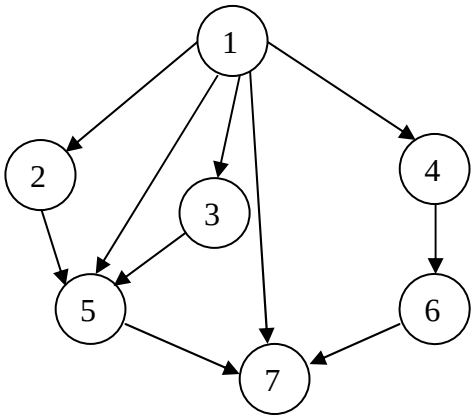

Figura 4.2.1: Representação de um Grafo Acíclico Direcionado (DAG).

As probabilidades condicionais podem ser extraídas a partir das equações 4.1.1 e 4.1.4. A construção de uma rede Bayesiana exige que certos cuidados sejam tomados de forma a permitir que a tabela conjunção de probabilidades resultante seja uma boa representação do problema.

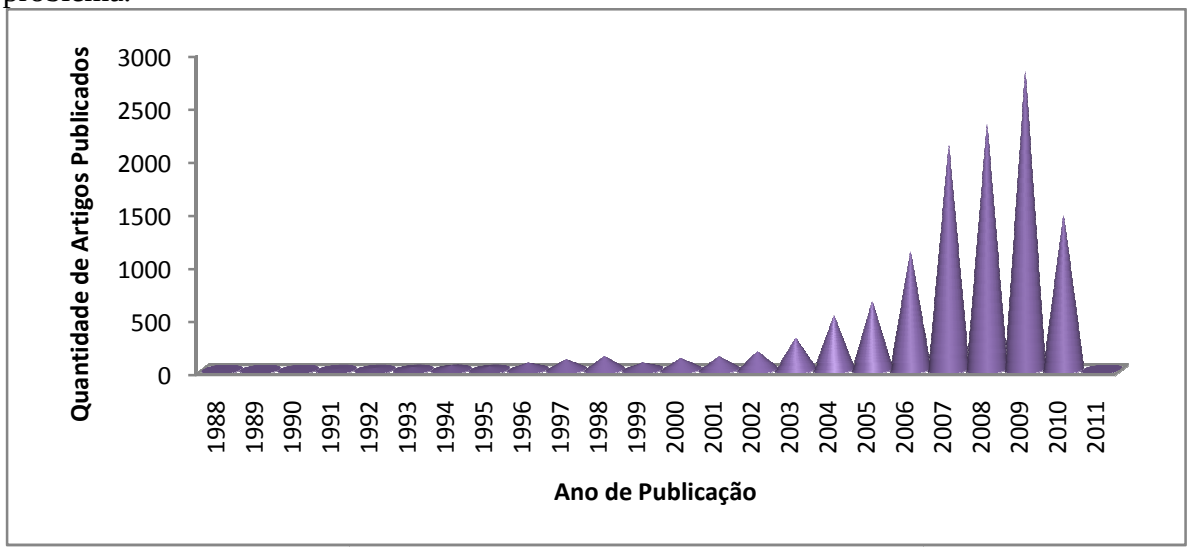

Figura 4.2.2: Número de artigos publicados ao ano sobre o tema Redes Bayesianas.

Ressaltamos ainda que o tema vem evoluindo desde o ano 1988, conforme ilustra a Figura 4.2.2, com um total de 12.134 publicações na área.

5. Aplicação em dados do Enem

O Exame Nacional do Ensino Médio (Enem) é uma proposta do Ministério da educação, criada em 1998 que, em seu histórico, objetiva avaliar o desempenho dos estudantes ao fim da escolaridade básica ou que já concluíram o ensino médio em anos anteriores.

O Enem é utilizado como critério de seleção para os estudantes que pretendem concorrer a uma bolsa no Programa Universidade para Todos (ProUni). Além disso, cerca de 500 universidades já usam o resultado do exame como critério de seleção para o ingresso no ensino superior, seja complementando ou substituindo o vestibular. Sendo assim, as universidades possuem autonomia e podem optar entre quatro possibilidades de utilização do novo exame como processo seletivo:

- Como fase única, com o sistema de seleção unificada, informatizado e on-line;
- Como primeira fase;
- Combinado com o vestibular da instituição;

- Como fase única para as vagas remanescentes do vestibular.

Atualmente, o Enem possui um objetivo maior que é a unificação nos processos seletivos das universidades públicas federais, motivada pelo que foi exposto na proposta à Associação Nacional os Dirigentes das Instituições Federais de Ensino Superior.

> Os exames de seleção para ingresso no ensino superior no Brasil (os vestibulares) são um instrumento de estabelecimento de mérito, para definição daqueles que terão direito a um recurso não disponível para todos (uma vaga específica em determinado curso superior). O reconhecimento, por parte da sociedade, de que os vestibulares são necessários, honestos, justos, imparciais e que diferenciam estudantes que apresentam conhecimentos, saberes, competências e habilidades consideradas importantes é a fonte de sua legitimidade. Parte-se aqui, portanto, do reconhecimento da necessidade, importância e legitimidade do vestibular. O que se quer discutir são os potenciais ganhos de um processo unificado de seleção, e a possibilidade concreta de que essa nova prova única acene para a reestruturação de currículos no ensino médio. (Enem, 2010)

Entretanto, para alcançar este objetivo com louvor é necessário acompanhamento planejado e estratégico que o Ministério da Educação vem propondo com a Matriz de Referência para o Enem em cada ano de sua execução. Porém, mais do que isso, são necessárias tecnologias e ferramentas apropriadas de tratamento de informação para geração de conhecimento que dê suporte à tomada de decisão. Baseado nessa afirmação pretende-se fazer um aparato da ferramenta sistema especialista probabilístico com método de representação do conhecimento em Redes Bayesianas para verificar o perfil dos estudantes de Santa Catarina, a partir de dados do questionário sócio-econômico dos candidatos ao Enem 2008, disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), em seu *site* na internet.

5.1. Conteúdo dos Dados e Arquitetura da Rede Bayesiana

Primeiramente, foi feito o download no site do Inep dos microdados referente ao ano de 2008 contendo aproximadamente 200 perguntas com as respectivas respostas dos candidatos ao Enem.

Foram selecionadas para a elaboração da rede Bayesiana, sete questões: Escolaridade do Pai, Escolaridade da Mãe, Renda Familiar, se o candidato ao Enem trabalhou durante o Ensino Médio, Período que cursou o Ensino Médio, Tipo de Instituição de Ensino que o candidato ao Enem estudou e a decisão que o candidato ao Enem irá tomar quando concluir o Ensino Médio.

Para a construção da Rede Bayesiana foi utilizado a *Shell* Netica devido ao fácil manuseio e vantagens que ela proporciona: gera uma apresentação gráfica de qualidade, que pode ser incorporada dentro de outros documentos; pode encontrar decisões ótimas para problemas de decisão seqüencial; soluciona diagramas de influência; etc. (NASSAR, 2007)

Conforme o item 4.2 deste artigo a Rede Bayesiana será composta pelos nós de entrada com suas respectivas evidências e o nó de saída constituída de hipóteses. No Quadro 5.1.1 está discriminado os nós que irão compor a rede e suas respectivas probabilidades frequentistas.

Quadro 5.1.1: Nós da Rede Bayesiana

Nós de ENTRADA da Rede	Evidências	Probabilidades
Escolaridade do Pai	Não estudo	≅ 2,22

	Da 1ª a 4ª serie do ensino fundamental	≅ 37,0
	⋮	⋮
Escolaridade da Mãe	Não estudo	≅ 2,31
	Da 1ª a 4ª serie do ensino fundamental	≅ 35,6
	⋮	⋮
Renda Familiar	Até 1 salário mínimo	≅ 6,93
	De 1 a 2 salários mínimos	≅ 29,9
	⋮	⋮
Trabalhou durante Ensino Médio	Sim, o tempo todo	≅ 24,3
	Sim, menos de 1 ano	≅ 31,1
	⋮	⋮
Período Cursado	Somente Diurno	≅ 41,5
	Maior parte no Diurno	≅ 13,8
	⋮	⋮
Tipo de Instituição de Ensino	Somente em escola pública	≅ 87,8
	A maior parte em escola pública	≅ 3,06
	⋮	⋮
Nó de SAÍDA da Rede	**Hipóteses**	**Probabilidades**
Decisão do Candidato	Já concluí o Ensino Médio	≅ 16,5
	Prestar vestibular e continuar os estudos no ensino superior	≅ 47,7
	⋮	⋮

A Rede esta arquitetada a partir do conhecimento SE saída (Decisão) ENTÃO entrada(s). As Figuras 5.1.1 e 5.1.2 ilustram a Arquitetura da Rede Bayesiana aplicado aos dados do Enem e as probabilidades condicionais da regra SE decisão ENTÃO escolaridade da Mãe, respectivamente.

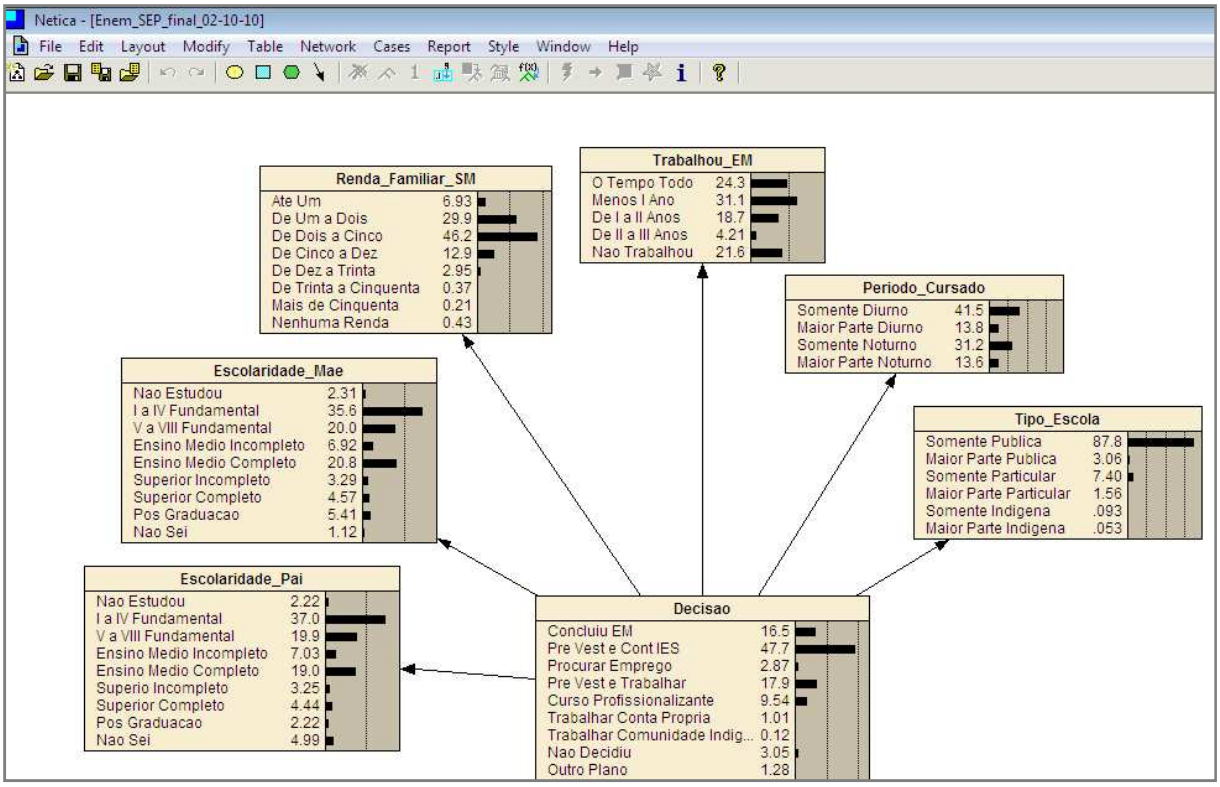

Figura 5.1.1: Arquitetura da Rede Bayesiana aplicada aos dados do Enem na *shell* Netica.

Node: Escolaridade_Mae

Chance | % Probability

Decisao	Nao Estudou	I a IV Fundame...	V a VIII Funda...
Concluiu EM	3.349	38.752	19.336
Pre Vest e Cont IES	1.631	30.937	19.41
Procurar Emprego	3.231	48.615	17.692
Pre Vest e Trabalhar	2.188	36.897	21.034
Curso Profissionalizante	3.55	44.512	22.513
Trabalhar Conta Propria	1.277	44.255	19.149
Trabalhar Comunidade I...	5.714	28.571	20
Nao Decidiu	3.184	42.692	20.839
Outro Plano	3.39	32.542	19.322

Figura 5.1.2: Probabilidades condicionais da regra SE decisão ENTÃO escolaridade da Mãe na *shell* Netica.

É válido salientar que quando selecionamos uma hipótese da saída (decisão) ou uma evidência de algum nó de entrada toda a rede se altera gerando probabilidades condizentes com a regra a qual queremos verificar para a tomada de decisão.

6. Resultados

Acionando algumas evidências (entradas) da rede obtemos como decisão (saída) que: se escolaridade do pai é igual a não estudou, escolaridade da mãe é igual a não estudou, renda familiar até um salário mínimo, candidato ao Enem trabalhou durante todo o ensino médio, cursou todo o ensino médio no período noturno e sempre estudou em escola pública, temos um resultante de 28,70% de chance de que o candidato almeje apenas a conclusão do ensino

médio em detrimento as outras hipóteses que tiveram um percentual inferior para este caso. Se das regras acima alterarmos apenas a evidência escolaridade da mãe para concluiu o ensino fundamental de primeira a quarta série, a saída da rede (decisão) se altera, resultando em um percentual maior para que o candidato almeje cursar um pré-vestibular e trabalhar concomitantemente.

Pode-se verificar que o perfil dos estudantes de Santa Catarina, segundo a classificação da rede Bayesiana, em geral, é marcada pela decisão dos candidatos em cursar pré-vestibular e continuar os estudos no ensino superior.

Verificou-se ainda que, alterando as probabilidades condicionais dos nós de entrada em até 20%, a rede continua classificando os dados sem que haja uma alteração dos resultados do nó de saída. Denominamos essa margem de “erro” como ruído da rede bayesiana.

Dessa forma, se tomarmos as mesmas variáveis para se fazer estimativas nos anos posteriores, nos dados do Enem, só haverá uma eventual mudança nas classificações, se as probabilidades condicionais ultrapassarem a margem de 20% no valor de seus dados.

7. Conclusões

A composição de Sistemas Especialistas Probabilísticos e Redes Bayesianas é uma excelente ferramenta de representação de conhecimento em um domínio específico, devido à habilidade para trabalhar ao nível do especialista; a representação em rede permite ao especialista expressar diretamente a relação qualitativa fundamental de "dependência direta" entre as variáveis do sistema; incorporar o processo de explanação e formas de manipulação de incerteza; entre outras peculiaridades.

A Rede Bayesiana gerada a partir dos dados do Enem sugere uma excelente alternativa de apoio à decisão para, por exemplo, auxiliar em políticas públicas direcionadas a educação ao que se refere principalmente no acesso e permanência às Universidades Brasileiras.

Uma complementação na rede através de novas inclusões de variáveis ou evidências e novas inferências pode ser realizada, na rede já estabelecida, conforme a necessidade dos especialistas para se obter maiores resultados no que se trata do avanço da proposta do ministério da Educação em relação ao acesso e permanência às Universidades através do Exame Nacional do Ensino Médio (Enem).

Referências

BRACARENSE e NASSAR, Silvia Modesto. Taxonomia da Ignorância. Disponível em: <*http://www.inf.ufsc.br/~silvia/*>. Acessado em 20 set. 2010.

CHARNIAK, Eugene. *Bayesians Networks without Tears*. IA Magazine, 1991. Exame Nacional do Ensino Médio (Enem). Disponível em <http://www.enem.inep.gov.br/enem.php>. Acessado em agosto de 2010.

FREITAS, Rodrigo E. A. *Um Portal para Cálculo de Probabilidades Geométricas*. Trabalho de Conclusão de Curso. Universidade Federal de Pernambuco. Nov. 2009, 33f.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Disponível em <http://www.inep.gov.br/>. Acessado em outubro de 2010.

KLIR, G.J. e FOLGER, T.A. *Fuzzy Sets, Uncertainty and Information*. Prentice Hall, 1998.

LIEBOWITZ, Jay. *Introduction to Expert Systems*. Santa Cruz, California, USA: Mitchell Publishing Inc., 1988

MARQUES, R. L.; DUTRA, I. *Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações*. Rio de Janeiro: [s.n.], 2008. Disponível em: <www.cos.ufrj.br/~ines-/courses/cos740/leila/cos740/Bayesianas.pdf>.

MLODINOW, Leonard. *O andar do bêbado: como o acaso determina nossas vidas*. Rio de Janeiro: Jorge Zahar Ed., 2009.

NASSAR, S. M. *A estatística como Apoio à Inteligência Artificial: Sistemas Especialistas Probabilísticos*. In: Estatística e Informática: um processo interativo entre duas ciências. Trabalho apresentado no Concurso para Professor Titular, INE, CTC, UFSC, 1998.

NASSAR, Silvia Modesto. *Tratamento De Incerteza: Sistemas Especialistas Probabilísticos.* Material Didático. Universidade Federal de Santa Catarina, 2007.

PANTALEÃO, Eliana. *Aplicação de técnicas de sistemas baseados em conhecimento em projeto cartográfico temático*. Dissertação de Mestrado, Universidade Federal do Paraná, Curitiba, 2003, 96 f.

PEARL, Judea. *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*. California, San Mateo: Morgan Kaufmann Publishers, 2 ed, 1988, 552 p.

RAMOS, R. F. *Sistemas especialistas - uma abordagem baseada em objetos com prototipagem de um selecionador de processos de soldagem*. Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianópolis, 1995.