Aplicação *de Data Mining* na Base de Dados do Processo Seletivo do Exame Nacional do Ensino Médio - ENEM 2010

Suely da Silva Carreira (FAP – Faculdade Adventista Paranaense) sscarreira@gmail.com
Manoel Francisco Carreira (Universidade Estadual de Maringá - UEM) mfcarreira@uem.br
Gilberto Clovis Antonelli (Universidade Estadual de Maringá - UEM) gcantonelli@uem.br
Márcia Marcondes Altimari Samed (Universidade Estadual de Maringá - UEM) mmasamed@uem.br
Gislaine Camila Lapasini Leal (Universidade Estadual de Maringá - UEM) gclleal@uem.br

Resumo:
A mineração de dados (MD) contribui para a otimização das informações e permitir que se possa gerar conhecimento para as tomadas de decisões e assim contribuir para o bom desempenho das organizações. Nesse aspecto, o objetivo deste trabalho é apresentar o processo de Descoberta do Conhecimento em banco de Dados (KDD), utilizando a mineração de dados, como uma ferramenta útil ao processo decisório de políticas públicas para área de ensino no país. Utilizado o banco de dados do Exame Nacional do Ensino Médio – ENEM-2010, tendo como ferramenta para o tratamento de dados o *software* "*WEKA*"- buscou-se nos dados do ENEM-2010 informações que possibilitaram a aquisição de conhecimento quanto as políticas públicas para a educação no Brasil. Os quesitos que nortearam o desenvolvimento do trabalho foi a verificar da existência de relação entre o desempenho do aluno, com o grau de escolaridade dos pais, acesso à internet e o tipo de escola que o aluno cursou (pública ou privada). A amostra é restrita aos alunos da Região Sul do país (Paraná, Santa Catarina e Rio Grande do Sul). A metodologia seguiu o padrão do processo KDD, seleção dos dados, pré-processamento, transformação, mineração de dados e interpretação. O banco de dados continha inicialmente por 4.200.000 alunos cadastrados, perfazendo cerca de 1,5 bilhões de dados. Os resultados confirmam conceitos já consolidados em termos da disponibilidade das tecnologias para aquisição do conhecimento. Assim a mineração de dados se mostrou uma ferramenta eficaz para a transformação de dados em conhecimento.

Palavras chave: Mineração de Dados, ENEM, KDD, Descoberta de Conhecimento.

Application of Data Mining in the Database of Selection Process Examination of National High School - 2010 ENEM

Abstract
Data mining (DM) contributes to the optimization of information and allow you to generate knowledge for decision-making and contribute to the performance of organizations. In this respect, the aim of this paper is to present the process of Knowledge Discovery in database (KDD), using data mining, as a useful tool for decision-making of public policy area of education in the country. Used the database of the National High School Examination - ENEM-2010, and as a tool for data processing software "WEKA" - sought on data ENEM-2010 information that enabled acquisition of knowledge as public policy for education in Brazil. The questions that guided the development of the work was checking the existence of a relationship between learner performance, with the level of parental education, internet access and the type of school the student attended (public or private). The sample is restricted to students from southern Brazil (Paraná, Santa Catarina and Rio Grande do Sul). The methodology followed the pattern of the KDD process, data selection, preprocessing, transformation, data mining and interpretation. The database contained 4.2 million initially for registered students, totaling about 1.5 billion data. The results confirm the concepts already established in terms of the availability of technologies for knowledge acquisition. Thus data mining has proved an effective tool for transforming data into knowledge.
Key-words: Data Mining, ENEM, KDD, Knowledge Discovery.

1. Introdução

No contexto atual, independentemente da região, país ou cidade onde uma organização desenvolva suas atividades, a necessidade de armazenamento de dados torna-se de suma importância para garantir aos seus gestores as informações referentes às operações por ela realizadas para serem utilizadas quando forem necessárias. O problema é que a quantidade de dados armazenados é cada vez maior, o que dificulta a geração de informações para a tomada de decisão.

Neste sentido, o presente trabalho, pretende apresentar a Mineração de Dados (MDs) como importante ferramenta de auxílio aos gestores públicos para a tomada de decisão, buscando mostrar que é possível "minerar" dados de um determinado banco de dados e transformá-los em informação para o processo decisório (políticas públicas) e assim, possibilitar a geração de conhecimento, que é o principal objetivo da MDs – para o caso em questão.

O Ministério da Educação e Cultura (MEC), anualmente, realiza o Exame Nacional do Ensino Médio (ENEM) com o objetivo de avaliar o desempenho dos estudantes do ensino médio para aferir o desenvolvimento de competências fundamentais ao exercício pleno da cidadania. Junto ao instrumento de exame são coletados dados socioeconômicos dos estudantes, o qual forma o banco de dados da avaliação deste trabalho. Em média, anualmente, participam do ENEM cerca de 5.000.000 de estudantes.

O foco da pesquisa busca associar o desempenho na prova objetiva com situações socioeconômicas, como o grau de escolaridade de seus pais, o acesso à internet e o tipo de escola em que o estudante cursou o ensino médio (pública ou privada). Os dados analisados são da Região Sul do País (estados do Rio Grande do Sul, Santa Catarina e Paraná).

2. Fundamentação teórica

2.1 Dados *versus* informações e conhecimento

Atualmente tem-se uma infinidade de dados apresentados através de redes de comunicação como jornais, revistas, televisão e principalmente a internet. Diariamente milhares de dados são "jogados - disponibilizados" para as pessoas, as quais muitas vezes não entendem o que estes dados significam e muitos se encarregam de retransmiti-los sem saber, na verdade, o que representam.

Os dados são apenas números que, isolados, não têm significado algum. A partir do momento em que esses dados são trabalhados e correlacionados com outros é possível ter informação. Os dados (as informações) são transformados em conhecimento à medida que for possível para aquele que o recebe tomar algum tipo de decisão ou mudar sua forma de pensar; ou seja, o conhecimento tende a proporcionar mudanças naquele que o adquire.

Pode-se dizer que a informação proporciona conhecimento quando para aquele que o recebe ocorre algum tipo mudança. Por exemplo, quando se perceber que algo pode ser realizado de maneira diferente, entender que não está realizando o procedimento correto ou da melhor forma. Assim, só se pode afirmar que tal informação gerou conhecimento se ela causar alguma reação ou mudança de comportamento.

O conhecimento tende a mostrar àquele que o adquire uma nova forma de ver o mundo, a fazer surgir uma nova vertente, um novo proceder, enquanto a informação, dependendo do seu nível, causa apenas um impacto momentâneo em quem o recebe, sem, na maioria das vezes, causar uma mudança comportamental.

2.2 Exame Nacional do Ensino Médio - ENEM

O ENEM foi instituído em 1998 pelo MEC para ser aplicado, em caráter voluntário, aos estudantes egressos do Ensino Médio (MEC 2012). Realizado anualmente, tem como objetivo principal avaliar o desempenho do aluno ao término do ensino médio, para aferir o desenvolvimento de competências fundamentais ao exercício da cidadania.

O exame tem como objetivo oferecer uma referência para que o estudante possa proceder à autoavaliação e escolher a competência profissional para a continuidade dos estudos. Busca estruturar, ao final da educação básica, uma avaliação que sirva como modalidade alternativa ou complementar aos processos de seleção nos diferentes mercados de trabalho, além de permitir acesso aos cursos profissionalizantes dos Pós-médios e à Educação Superior. Também credencia os estudantes à participação dos programas governamentais.

O ENEM consiste de uma prova única contendo 180 questões objetivas de múltipla escolha e uma proposta para redação, além de duzentas questões de natureza socioeconômica.

2.3 Descoberta de Conhecimento em Banco de Dados

O KDD (*Knowledge Discovery in Databases*) é um processo não trivial de identificação de padrões. De acordo com Fayyad *et al*. (1996), esse processo deve conter na base de dados as características de validade, novidade, utilidade e assimilabilidade. O KDD é o processo de selecionar e processar dados que permitam identificar estruturas interessantes que possam extrair conhecimento dos dados, e para isto aplica-se a mineração de dados. A expressão Mineração de Dados (MDs) refere-se a uma das etapas deste processo.

2.3.1 Mineração de Dados - MDs

A MDs consiste em abstrair de um banco de dados informações que gerem conhecimento e possam auxiliar no processo de tomada de decisão. De acordo com Fayyad *et al*. (1996), a mineração de dados é a principal etapa do processo KDD, e está voltada a aplicar algoritmos e produzir padrões sobre uma base de dados.

A MDs de acordo com a DWBrasil (2004), segue três caminhos. O primeiro deles é a estatística clássica, que envolve conceitos básicos (distribuição normal, variância, etc.) usados para estudar os dados e os relacionamentos entre eles; o segundo caminho traçado pela MDs é a Inteligência Artificial, a qual é construída a partir dos fundamentos da heurística, em oposição à estatística, e tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos; e o terceiro caminho é a aprendizagem de máquina (*machine learning*), que pode ser compreendida como a junção entre a estatística e a Inteligência Artificial. A aprendizagem de máquina tenta fazer com que os programas de computador aprendam com os dados que utilizam, de tal modo que esses programas tomem decisões diferentes, baseadas nas características dos dados, usando a estatística para os conceitos fundamentais e adicionando heurística avançada da Inteligência Artificial e algoritmos.

O uso da MDs como parte do processo KDD tem grande potencial para auxiliar as organizações na extração de informações provenientes dos seus bancos de dados, predizendo padrões e comportamentos futuros e respondendo a questões que tomariam muito tempo para serem resolvidas, o que possibilita tomar decisões corretas, por estarem apoiadas em conhecimento. A MDs dispõe de tarefas básicas classificadas nas categorias descritivas e preditivas, entre as quais se podem citar: a classificação, associação, segmentação (ou *clustering*), estimativa (ou regressão) e sumarização. A seguir serão descritas as tarefas de MDs utilizadas para a descoberta do conhecimento neste trabalho.

2.3.2 Classificação – Associação – Segmentação – Árvore de Decisão e WEKA

A tarefa de classificação consiste em construir um modelo que possa ser aplicado a dados não classificados com vista a dividi-los em classes. Os dados são analisados e separados por classes. A classificação tem como resultado a construção da "árvore de decisão", que apresenta uma visualização gráfica das diferentes correlações dos dados, permitindo estabelecer a classificação de cada elemento. A tarefa de classificação pode ser considerada uma tarefa maldefinida (indeterminística), que é inevitável, em caso, em que envolve predição, como os processos de classificação dentro do contexto da Mineração de dados (Freitas, 2000, p.65).

A associação consiste em determinar quais itens estão correlacionados, ou seja, costumam ser encontrados juntos nos mesmos tipos de transação. A tarefa de associação é considerada bem determinística e não envolve processo de predição como resultado final.

Segmentar significa dividir grupos heterogêneos em subgrupos mais homogêneos. Na segmentação de dados não há classes predefinidas, os registros são agrupados de acordo com a semelhança, formando subgrupos que permitem a visualização de relações de analogia.

A árvore de decisão é uma representação gráfica de alternativas disponíveis geradas a partir de uma decisão inicial que pode servir de apoio às tomadas de decisão. Uma das grandes vantagens de uma árvore de decisão é a possibilidade de transformação ou de decomposição de um problema complexo em diversos subproblemas mais simples.

Para efetuar a representação gráfica da árvore de decisão são geralmente usadas linhas para identificar a decisão (por exemplo, "sim" ou "não") e nós para identificar as questões sobre as quais se deve decidir. Cada um dos ramos formados por linhas e nós termina numa espécie de folha que identifica a consequência mais provável da sequência de decisões.

O WEKA (*Waikato Environment for Knowledge Analysis*) é uma ferramenta que permite realizar a MDs. Consiste num software utilizado na linguagem Java e desenvolvido no meio acadêmico da Universidade de Waikato, na Nova Zelândia, em 1999. Tem como vantagem o fato de ser de domínio público. É uma ferramenta formada por um conjunto de algoritmos que implementam diversas técnicas para resolver problemas de MDs.

3. Metodologia

O presente trabalho realizou a MDs do banco de dados do Enem 2010. Os dados minerados são dos estados de Rio Grande do Sul, Santa Catarina e Paraná. O banco de dados é disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), vinculado ao Ministério da Educação (MEC) inicialmente em arquivo-texto (txt) de aproximadamente 5 GB.

Para acessar o banco de dados foi utilizado o software IBM-SPSS, software estatístico que permite extrair de um determinado banco de dados (relativamente grande) somente os dados que interessa para o foco da pesquisa.

No banco de dados do Enem 2010 existem registrados aproximadamente 4.200.000 (quatro milhões e duzentos mil) alunos que se inscreveram em todo o Brasil para prestar o exame. As questões socioeconômicas a serem respondidas são aproximadamente 200 (duzentas), assim o número de informações existentes neste arquivo é aproximadamente 1.500.000.000 (um bilhão e meio) Após a utilização do software estatístico o arquivo SPSS apresentou um tamanho aproximado de 10 GB.

A partir disso foi realizada a fase de pré-processamento, em que foram selecionados somente os dados dos estados de Santa Catarina, Rio Grande do Sul e Paraná. Neste contexto foram selecionadas três questões (das duzentas) e o arquivo foi reduzido para 500 MB.

Na seqüência foram eliminados do banco de dados os alunos que não compareceram à prova, o que resultou em aproximadamente 370.000 (trezentos e setenta mil) alunos dos três estados. O arquivo gerado em planilha eletrônica apresentou um tamanho de 30 MB.

As questões escolhidas para o estudo foram as seguintes:

Questão "A" – "Até quando seu pai estudou?", alternativas de resposta: analfabeto "ANA", 1º grau "1GR", 2º grau "2GR" e superior "SUP". Inicialmente, também se selecionou a questão que indagava a escolaridade da mãe, porém o resultado se mostrou similar ao primeiro. Optou-se, então, por utilizar apenas a escolaridade relativa ao pai.

Questão "B" – "Você tem Acesso à Internet", alternativas de resposta: não tem acesso "NTI" e tem acesso "TAI".

Questão "C" – "Que tipo de escola você cursou ou está cursando o ensino médio (2º grau)?", alternativas de resposta: escola pública "PUB" e escola privada "PAR".

E o resultado da nota da prova objetiva, as quais foram transformadas em conceitos, da seguinte forma:

Quadro 1 – Relação de transformação nota em conceito – prova objetiva

Nota da prova objetiva	Atribuição de conceito – prova objetiva
0,00 a 3,99	Insatisfatório – "INS"
4,00 a 5,99	Regular – "REG"
6,00 a 7,99	Bom – "BOM"
8,00 a 10,0	Excelente – "EXC"

As questões que motivaram o desenvolvimento do trabalho foram:

a) O valor do desempenho (conceito – prova objetiva) do aluno tem relação com o grau de escolaridade dos pais?
b) Os alunos que têm acesso à internet obtiveram melhor desempenho (conceito – prova objetiva) no exame?
c) O aluno cursar escola pública ou privada interfere no seu desempenho no ENEM?

Através de resposta deste conjunto mínimo de questões é possível analisar a efetividade das políticas de governo.

4. Contextualização dos dados do ENEM 2010

4.1Dados gerais

O universo amostral é referente aos três estados da Região Sul do País para os quais constam 324.240 (trezentos e vinte quatro mil, trezentos e quarenta) alunos que responderam o questionário socioeconômico validado. Destes 46,5% são do Paraná, 13,0% de Santa Catarina e 40,5% do Rio Grande do Sul. A Tabela 1 apresenta o número de alunos e o percentual de cada uma das três questões, assim como o resultado da padronização do conceito obtido pela correção das questões objetivas de conteúdo geral e específico.

Tabela 1 – Dados referentes aos estados Paraná, Santa Catarina e Rio Grande do Sul.

Questão	Opções	Paraná 150.633		Santa Catarina 42.344		Rio Grande do Sul 131.263	
		Nr. de alunos	%	Nr. de alunos	%	Nr. de alunos	%
Que tipo de escola que cursou 2º grau?	Escola pública	132.240	88,0	36.314	86,0	119.638	91,0
	Escola privada	18.393	12,0	6.030	14,0	11.625	9,0
Tem acesso a internet?	Acesso Internet	75.996	50,5	22.768	54,0	52.883	40,0
	Não tem acesso	74.637	49,5	19.576	46,0	78.380	60,0
Escolaridade dos pais?	Analfabetos	4.506	5,0	974	2,0	4.506	3,5
	1º grau	83.706	55,5	23.131	55,0	51.921	40,0
	2º grau	39.196	26,0	10.932	26,0	61.346	47,0
	Superior	19.760	13,5	7.307	17,0	13.490	10,0
Conceito na prova objetiva?	Insatisfatório	71.189	47,0	20.051	47,0	53.521	41,0
	Regular	57.166	38,0	16.949	40,0	57.189	43,5
	Bom	20.186	13,5	4.792	11,0	19.359	15,0
	Excelente	2.092	1,5	552	2,0	1.194	0,5

Pelos dados constata-se que, em média, entre 88 e 91% dos alunos do ensino médio cursaram escola pública e menos de 10% cursaram em escola privada. Quanto à escolaridade dos pais, nos estados do Paraná e Santa Catarina, cerca de 55% cursaram apenas o primeiro grau, enquanto no Rio Grande do Sul o percentual é de apenas 40%. Entretanto, para os pais com escolaridade de ensino médio no Rio Grande do Sul têm-se 47%, contra 26% nos estados de Santa Catarina e Paraná. O percentual de pais com escolaridade de nível superior está entre 10 e 14% em relação aos três estados.

Considerando a Tabela 1, perceber-se que para o mesmo quesito os dados dos três estados convergem para uma mesma faixa de valores, que é representado pelos gráficos das Figuras 1 e 2 (estados do Paraná e de Santa Catarina).

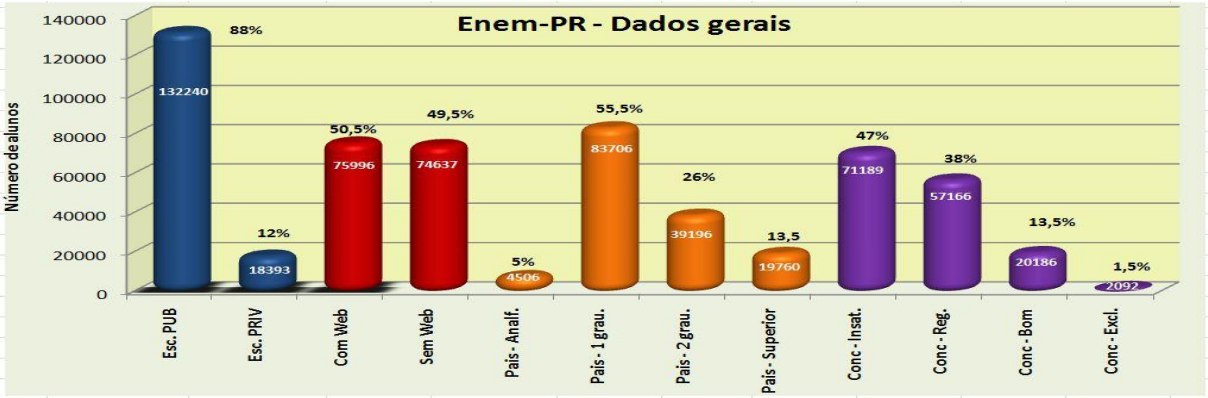


Figura 1 – Dados gerais – ENEM – PR (2010).

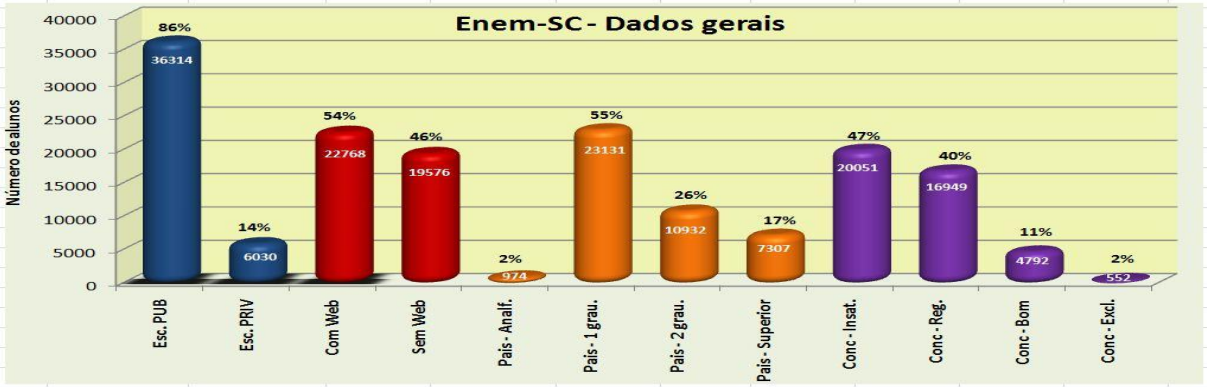


Figura 2 – Dados gerais – ENEM – SC (2010).

Quanto ao acesso a internet, o resultado demonstra um percentual semelhante entre os alunos com acesso e sem acesso para os estados do Paraná e Santa Catarina. A diferença surge no Rio grande do Sul, no qual os alunos sem acesso à web representam cerca de 60%.

Em relação aos conceitos nos três estados, em média, 85% dos alunos obtiveram conceito insatisfatório e regular, ou seja, notas abaixo de 6,0 (seis). Deve-se salientar que a média abaixo de 6,0 (seis) significa que uma grande parte dos alunos apresentou nível de conhecimento inferior a 60% do que foi exigido como conhecimento máximo. O resultado indica a necessidade de melhorar a qualidade do ensino e rever metodologias e práticas educacionais vigentes.

4.2 Resultado do processo de classificação – árvore de decisão

Utilizando o software "Weka", testou-se o processo de classificação tendo como base de análise os quatro quesitos de investigação de trabalho, conseguiu-se a definição de árvores de decisão apenas por meio do algoritmo "J48"; porém, apesar da geração de diversas árvores, a maioria delas não atendeu ao requisito de confiabilidade (valor de Kappa - medida de concordância usada em escalas nominais que nos fornece uma idéia de quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando-nos assim quão legítimas são as interpretações) baixo e erros absolutos elevados) dos resultados.

Quando se tem como base a árvore de decisão do desempenho "conceito" da prova objetiva em relação às demais questões, observa-se que a geração da árvore de decisão (ex., Figura 3) apresenta o índice "kappa" muito baixo (SC→K=0,17 PR→K=0,13 e RS→K=0,1) e não representativo (concordâncias "pobres"). Em relação ao percentual de erro de predição, os valores de acerto também foram baixos, entre 45 e 55%.

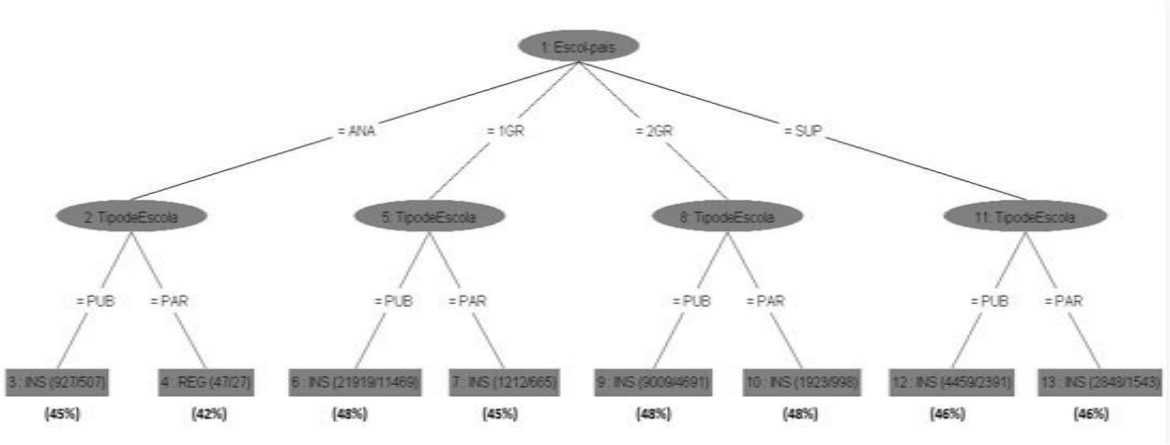


Figura 3 – Árvore de decisão – desempenho na prova – estado Santa Catarina.

Percebe-se claramente, pela Figura 3, que não se obteve classificação em função do desempenho na prova objetiva, o que se deve ao fato da maioria absoluta dos alunos ter obtido conceitos insatisfatório ou regular, independentemente da divisão dos demais quesitos (o conceito insatisfatório se mostrou distribuído em todos os quesitos com percentual oscilando entre 40 e 50%). Tal situação se reproduziu nos três estados pesquisados, conforme demonstram os índices de "Kappa".

No quesito de acessibilidade à internet o índice de "Kappa" se mostrou melhor que o relativo ao "desempenho" dos alunos na prova objetiva. Os valores de "Kappa" são os seguintes: SC→K=0,29, PR→K=0,30 e RS→K=0,25, os quais apresentam classificação "razoáveis". Em relação aos erros absolutos de predição, a faixa de acerto oscilou entre 65 e 74%, valores que se classifica como "razoáveis".

Exemplo típico deste resultado é visualizado na Figura 4, cujos dados se referem ao Estado de Santa Catarina tendo como a base da árvore o acesso à internet.

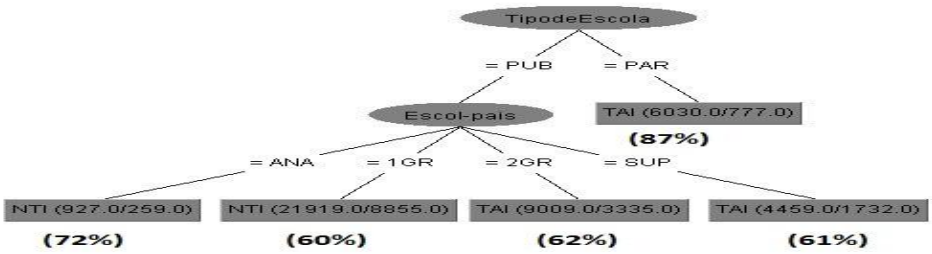


Figura 4 – Arvore de decisão – base acesso internet – estado de Santa Catarina

No caso da árvore apresentada na Figura 4, os dados permitem colocar no primeiro nível os alunos de escola privada (particular), os quais tem acesso à internet; já para os que frequentam escola pública, os com acesso à internet somente obtém classificação no segundo nível. Para o item escolaridade dos pais, ou seja, aluno de escola pública cujos pais tenham escolaridade de 2º grau ou superior os mesmo em sua maioria tem acesso à internet, enquanto os alunos cujos pais são analfabetos ou tem apenas o 1º grau, os mesmos em sua maioria não têm acesso à internet. Percebe-se que quando o grau de escolaridade dos pais é mais elevado o aluno tem acesso à internet, realidade que tem como um dos prováveis fatores a renda familiar, que em geral é mais elevada, devido ao fato do grau de escolaridade ser maior. Outro fator é em geral à necessidade dos pais terem internet para poderem desenvolver suas atividades profissionais. Ainda, em relação ao acesso à internet, o Estado do Rio Grande de Sul, com o índice "kappa" classificado como "razoável" e acerto de predição superior a 65%, apresentou uma árvore distinta em relação aos estados de Santa Catarina e Paraná, como pode ser visto na Figura 5.

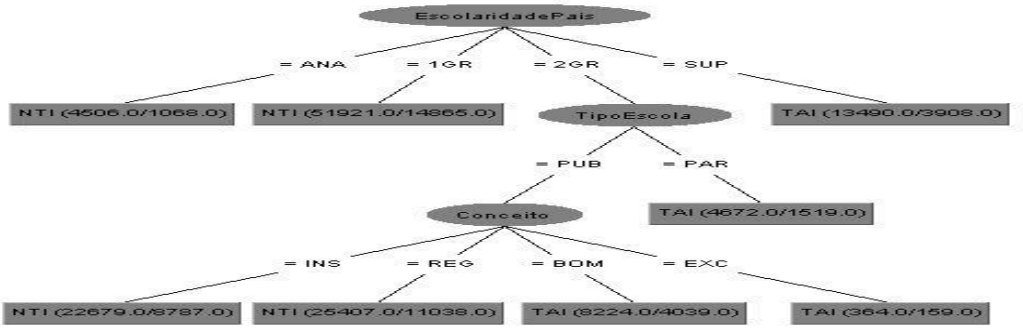


Figura 5 – Árvore de decisão – base acesso internet – estado de Rio Grande do Sul.

A árvore com base no acesso à internet para o Estado do Rio Grande do Sul (Figura 5) diferenciou-se de Santa Catarina (Figura 4), pois apresentou três níveis de classificação. No primeiro, a partir da escolaridade dos pais se classificou mais de 50% da amostra, pois para os graus de escolaridade "analfabeto", "1º grau" e "superior" finalizou a classificação de árvore, porém em relação aos alunos cujos pais têm escolaridade de 2º grau foram necessários o segundo e terceiro níveis de classificação para a decisão final. No segundo nível fez-se uso dos dados em relação ao tipo de escola; neste caso, quanto aos alunos de escolas privadas foi finalizada a classificação com "ter acesso à internet", porém para os alunos das escolas públicas a classificação exigiu mais um nível.

O quesito analisado foi o conceito da prova objetiva, ou seja, os conceitos insatisfatório e regular geraram a classificação do aluno como não tendo acesso à internet e os conceitos "bom" e "excelente" se referiram ao aluno que tem acesso à internet.

Em relação ao aluno a ter cursado o 2º grau (ensino médio) em escola pública ou privada, a referência da questão é se o tipo de escola em que o aluno concluiu o 2º grau tem relação com o desempenho "conceito" obtido na prova objetiva. Em parte, a análise do desempenho ficou comprometida em vista do percentual excessivo de notas com os conceitos insatisfatório e regular, que nos três estados totalizam mais de 80%, de forma distribuída em todos os quesitos. Tal situação foi confirmada pelo índice "Kappa", que para o Estado do Rio Grande do Sul apresentou valores praticamente zero, porém para os estados do Paraná e Santa Catarina os valores oscilaram entre 34 e 36% (considerada uma concordância "moderada") e em relação ao acerto dos dados de predição oscilaram entre 89 e 91%.

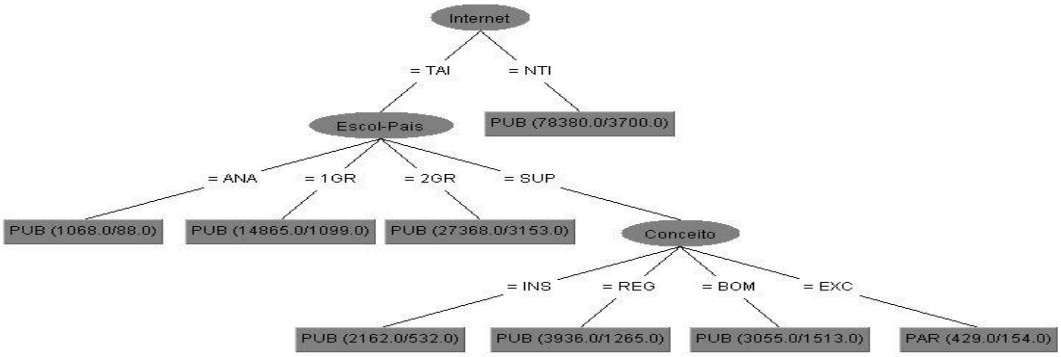


Figura 06 – Arvore de decisão – base escolaridade dos pais – Estado do Rio Grande do Sul

Considerando que para o Estado do Rio Grande do Sul a árvore de decisão (Figura 6) apresenta confiabilidade de classificação pelo índice "Kappa" e acerto de predição, pode-se observar que a classificação está distribuída em três níveis. No primeiro, não ter acesso à internet, resulta na classificação dos alunos que cursaram o 2º grau em escola pública. Para os que têm acesso à internet a árvore requer um segundo nível, utilizando a questão "escolaridade dos pais" assim utilizou-se a classificação para alunos em escola pública e com pais com escolaridade "analfabeto", "1º grau" e "2º grau", enquanto para os alunos que tem pais com escolaridade "superior" a classificação exige o terceiro nível que faz uso da questão referente ao desempenho dos alunos na prova objetiva (conceito), que define que os alunos com conceito "excelente" cursou 2º grau em escola particular. Para os demais conceitos os alunos foram classificados com estudantes de escola pública.

Em relação ao Estado do Paraná, com índice "Kappa" igual a 0,36 (concordância "moderada"), a árvore gerada (Figura 7) se mostrou distinta da gerada em relação ao Estado do Rio Grande do Sul (Figura 6).

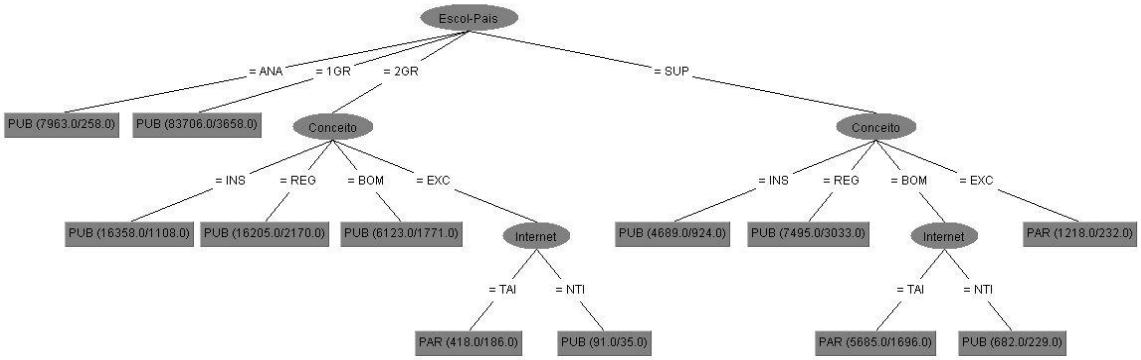


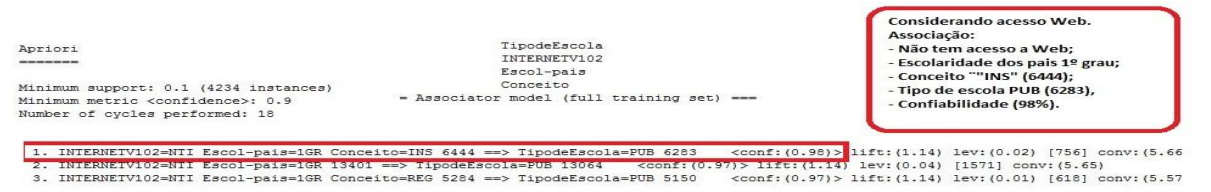
Figura 7 – Arvore de decisão – base escolaridade dos pais – Estado do Paraná

Observa-se pela Figura 7 que, para definir o tipo de escola em que aluno cursou o 2º grau foram exigidos três níveis para concluir a classificação. No primeiro nível foram classificados como estudantes de escola pública aqueles cujos pais são "analfabetos" ou têm "1º grau". Para os estudantes cujos pais têm 2º grau e/ou curso superior, a classificação exigiu a informação sobre o desempenho "conceito". Dentre os alunos cujos pais têm 2º grau incluíram-se na classificação os estudante de escola pública os que obtiveram conceitos "insatisfatório", "regular" e "bom"; para o conceito "excelente" se exigiu o terceiro nível. Em relação ao segundo nível, os estudantes com pais com curso "superior" foram classificados como concluintes de 2º grau em escola pública aqueles com desempenho "insatisfatório" e "regular", enquanto os com conceito "excelente" foram classificados como alunos de escola particular. Em relação aos alunos filhos de pais com escolaridade de 2º grau, se exigiu o terceiro nível. Neste último nível relacionou o acesso à internet.

4.3 Resultado do processo de associação

No processo de associação ou correlação, baseado no algoritmo "Apriori", os resultados encontrados foram semelhantes para os estados de Santa Catarina e Paraná (Figura 8) e diferentes para o Estado do Rio Grande do Sul (Figura 9).

Dados de Santa Catarina



Dados do Paraná

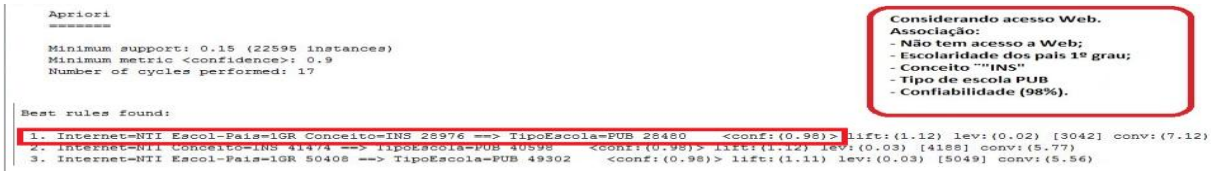


Figura 8 – Dados de associação – estados de Santa Catarina e Paraná

Pela Figura 8 constatou-se a seguinte associação: o estudante não ter acesso à internet, os pais com escolaridade de "1º grau", desempenho da prova objetivo com conceito "insatisfatório" e estudar em escola pública. Tal associação tem confiabilidade de 98%. Para o Estado do Rio Grande do Sul a associação pode ser visualizada na Figura 9.

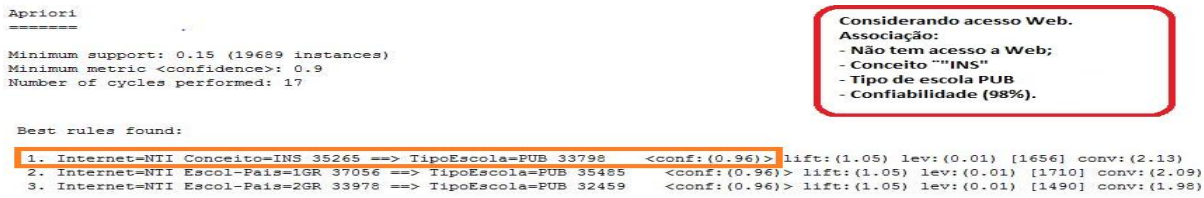
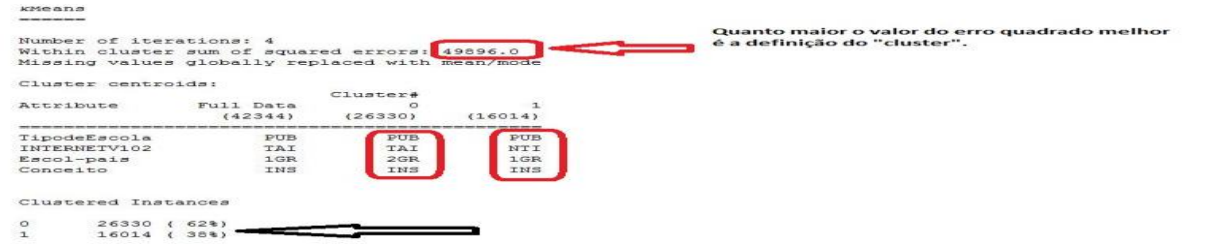


Figura 9 – Dados de associação – Estado do Rio Grande do Sul

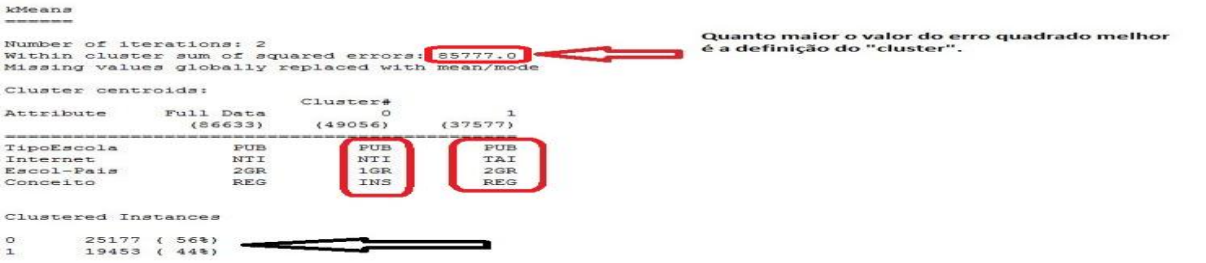
A diferença existente entre as Figuras 8 e 9 é que na associação gerada para o Estado do Rio Grande do Sul a questão referente à escolaridade dos pais não faz parte da associação. Excluída a escolaridade dos pais nesta associação, os restantes das combinações permaneceram inalteradas, alterando-se apenas o percentual de confiança de 98 para 96%.

4.4 Resultado do processo de segmentação (*clustering*)

O processo de clustering apresentou grupos distintos para os três estados, porém uma análise sucinta permite visualizar que os grupos não diferem significativamente. A figura 10 mostra os resultados da clustering por estado, em ordem crescente do valor do erro quadrado (quanto maior o valor melhor o *clustering*). Com base neste princípio a maior consistência está no grupo formado pelos dados do Estado do Paraná, percebendo-se que quanto maior o número de dados melhor é a consistência.



Dados do Estado do Rio Grande do Sul



Dados do Estado do Paraná

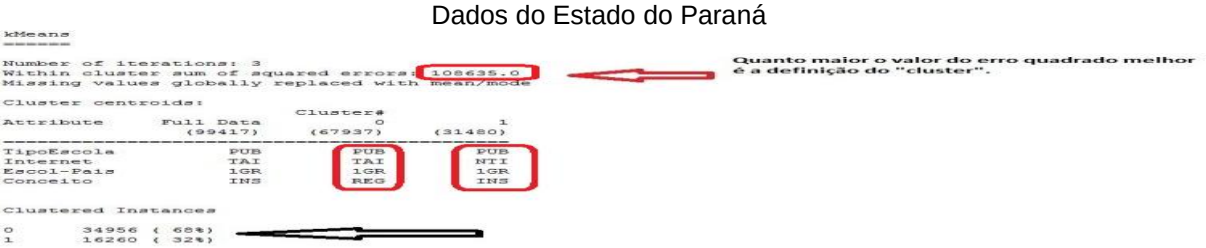


Figura 10 – Formação de clusters – para estados da região Sul

O melhor *cluster* formado foi o baseado nos dados do Estado do Paraná, e nele foram agrupados estudantes de "escola pública" que "têm acesso à internet", cujos pais têm escolaridade de "1º grau" e que obtiveram desempenho "regular" na prova objetiva; já o segundo agrupamento altera o primeiro da seguinte forma: o mesmo estudante "que não tem acesso à internet" obteve desempenho "insatisfatório" na prova de questões objetivas, o que mostra lógica dentro do contexto de aquisição de conhecimento atualmente existente.

Os resultados apresentados neste trabalho têm três dimensões: a primeira refere-se à classificação dos dados através da geração de árvores de decisão; a segunda, pela formação de associação baseada em coeficiente de confiabilidade; e a última, pela formação de clusters, que tem como principio a formação de agrupamento por analogia.

Os resultados obtidos no processo de classificação devem ser usados com cautela, por se tratar de predição. Percebeu-se que a consistência dos resultados deixou a desejar, em função dos índices "Kappa", assim como os percentuais de acerto se mostram modestos. Porém, apesar, da relativa inconsistência, podem-se perceber fatos interessantes, antes inquestionáveis como, não ser possível afirmar que o acesso à internet melhore o desempenho dos estudantes na prova de conhecimento. Nem que o grau de escolaridade dos pais e o tipo de escola em que o aluno cursou o 2º grau interfiram significativamente neste desempenho.

Em relação ao processo de associação, por não se tratar de predição, e sim, de constatação, foram obtidos dados de maior consistência, muitos deles com confiabilidade superior a 98%. Isso traduz uma lógica aceitável, como por exemplo, o fato de "não ter acesso à internet", ter "pais" com escolaridade de "1º grau", apresentar desempenho "insatisfatório" na prova de conhecimento, o aluno deve estudar em "escola pública". Tal perfil é perfeitamente coerente com nossa realidade educacional.

Para a *clustering* a formação de grupos análogos mostrou forte consistência e o resultado demonstrou tal fato, tendo o melhor grupo o seguinte perfil: estudantes de "escola pública", "com acesso à internet", filho de pais com escolaridade de "1º grau" obtiveram desempenho "regular" na prova de conhecimento. No segundo agrupamento, o "não acesso à internet" resultou em desempenho "insatisfatório" na prova de conhecimento, fato perfeitamente justificável dentro do nosso contexto de ensino do país.

Dos resultados obtidos pode-se concluir que a utilização da mineração de dados (*data mining*) permite extrair dos dados informações relevantes para a construção do conhecimento. Tal fato ocorreu no tratamento dos dados referentes ao Enem 2010, em que pela MDs foi possível, em alguns casos, constatar certas máximas apregoadas de forma intuitiva, e em outros, desmistificar conceitos fortemente estruturados.

Inicialmente tinha-se a ideia de que o fato de o estudante ter acesso a tecnologias que permitem buscar informações das mais diversas áreas e em qualquer parte do mundo poderia proporcionar melhor desempenho nas provas do ENEM; porém o trabalho demonstrou que não podemos afirmar que ter acesso à informação permite melhor desempenho.

Pode-se supor que os estudantes acessem a internet como meio de interagir com outras pessoas, utilizando ferramentas como e-mail, *Orkut, mensager, facebook* e outros, e não utilizem a tecnologia para pesquisas e estudos. O trabalho não permite fazer está afirmação, mas apenas fazer uma suposição que indica um novo foco de trabalho.

Referências

CABRAL JUNIOR, José Edilson et al**.** *Paradigma simbólico de aprendizagem aplicado ao banco de dados do vestibular da UFMS*. 2002. Disp. em: www.dct.ufms.br/~mzanusso/producao/JedRodrRog.pdf. 10/09/ 2010.
CARVALHO, D. R.. *Data Mining através de introdução de regras e algoritmos genéticos*, 1999. Dissertação para obtenção do grau de Mestre – PUCPR, Curitiba, 1999.
CARVALHO, D. R. *Um método híbrido Árvore de Decisão / Algoritmo Genético para Data Mining*, 2002. Tese realizada para obtenção do título de Doutor – PUCPR, Curitiba, 2002.
DIAS, M. M. *Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados*. 2001.Tese de Doutorado do PPGEP-UFSC. Florianópolis, SC.
DWBrasil Disponível em: http://www.dwbrasil.com.br/html/dmining.html Acesso em: 27 Jun. 2004.
FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. From *data mining to knowledge discovery: an overview*. In: *Advances in knowledge discovery and Data Mining*, AAAI Press / The MIT Press, MIT, Cambridge, Massachusets, and London, England, 1996, p.1-34.
FREITAS JUNIOR, OLIVAL G. *et al.* *Sistema de apoio à decisão usando a tecnologia Data Mining com estudo de caso da Universidade Estadual de Maringá*. I Congresso Brasileiro de Computação. Maringá, 2001.
GARCIA, S. C. *O Uso de árvore de decisão na descoberta de conhecimento na área da saúde*. 2000. Disponível em: www.inf.ufrgs.br/pos/SemanaAcadêmica/Semana 2000/Simone Garcia. Acesso 12/08/2010.
HARRISON, T. H, "*Intranet data Warehouse*", Editora Berkeley, 1998
INEP. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*. Disponível em: http://www.inep.gov.br/imprensa/noticias/censo/superior/news 04-05-imp.htm. Acesso em 07 de agosto de 2011.
***WEKA*.** Disponível em http://www.cs.waikato.ac.nz/~ml/weka/, 2010.
SPSS. *Data Mining*: *An Introduction*. Disp. http://www.spss.com/datamine/index.htm Acesso: 12 Jul. 2011.
SPSS. *Data mining and statistics - Gain a competitive advantage*. Disponível em: http://www.spss.com/datamine/index.htm Acesso em: 12 Jul. 2011.