Prática de Mineração de Dados no Exame Nacional do Ensino Médio

Leandro A. Silva, Anderson Hideki Morino, Thiago Massahiro Conti Sato

Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie
Rua da Consolação, xx – São Paulo –SP – Brasil

leandroaugusto.silva@mackenzie.br,
anderson_morino@hotmail.com,thiago.nbj@gmail.com

Abstract. *This paper presents a study of Educational Data Mining. More specifically, we use a task known as Data Association to find patterns of rules in test scores and socioeconomic questionnaires of the National Secondary Education Examination (ENADE) Data. The different parameterizations of the association algorithm known as Apriori allows to discover problems in national education as students studying at public school in general has regular performance on the exam; parents with studies of first degree have children studying in public school and regular performance on the exam; and as a last important example of a result, students with regular performance, family with 3 people and residents in the São Paulo study in public school .*

Resumo. *Este artigo apresenta um estudo de Mineração em Dados Educacionais. Mais especificamente, utilizamos uma tarefa da Mineração de Dados conhecida por Associação de Dados para encontrar padrões de regras nos resultados de provas e questionários socioeconômicos do Exame Nacional de Ensino Médio (ENADE) de Dados. As diferentes parametrizações do algoritmo de associação A Priori permite descobrir problemas na educação nacional como alunos que estudam em escola publica em geral tem desempenho regular no exame; pais com estudo do primeiro grau tem filhos que estudam em escola publica e com desempenho regular no exame; e como último importante exemplo de resultado, alunos com desempenho regular, família com 3 pessoas e moradores na região de São Paulo estudam em escola publica.*

1. Introdução

A educação é um setor fundamental para o crescimento e desenvolvimento de um país. A avaliação atual do Brasil indica a necessidade de melhorias neste setor, as quais poderiam ser conseguidas através do auxílio da Tecnologia de Informação [Santos, 2012].

Anualmente, a qualidade do ensino médio brasileiro é avaliada através do ENEM (Exame Nacional do Ensino Médio) e, desde 2009, tornou-se critério adicional de ingresso nas principais universidades do Brasil. Portanto, os objetivos principais do ENEM são: avaliar o ensino médio e gerar um indicador de vestibular para ingressar nas principais universidades do Brasil. E assim sendo, este exame se torna ainda mais importante para o crescimento do pais.

O ENEM trata-se de uma prova aplicada individualmente em diversas instituições situadas em todo o território nacional, na qual participam milhões de

estudantes, obtendo-se um resultado que varia numa nota de 0 a 1000. Além da prova, há um questionário socioeconômico cultural, com preenchimento não obrigatório ao aluno, a ser respondido no dia do exame.

Os dados de resultado das provas e do questionário socioeconômico cultural respondido podem se tornar fontes de dados para extrair diversas informações relevantes que possam ser importantes às Instituições de Ensino para tomada de decisões, investimentos, planejamentos estratégicos, entre outros. Neste processo de transformação de dados em informação insere-se à Mineração de Dados.

A Mineração de Dados é um passo no processo de Descoberta de Conhecimento em Base de Dados ou KDD (do inglês, *Knowledge Descovery in Database*) que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados [FAYYAD, 1996].

Mineração de Dados Educacionais ou EDM (Educational Datamining) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados oriundos de contextos educacionais [Romero & Ventura,2010], [Paiva et al., 2012]. Os tipos de estudos desta área são classificados, segundo Romero & Ventura (2010) em: educação *offline* para analises em dados de desempenho do aluno, comportamento, currículo e etc, ou seja, gerados em ambientes de sala de aula; aprendizado eletrônico (*e-learning*) e Sistema de Gestão da Aprendizagem (ou LMS, do inglês, *Learning Management System*) para analise de dados armazenados em sistemas LMS no formato de log e bases de dados; e Sistemas Tutores Inteligentes (ou ITS do inglês *Intelligent Tutoring System*) e Sistemas Hipermídias Adaptativos Educacionais (ou AEHS, do inglês, *Adaptive Educational Hypermedia System*) os quais são dados de sistemas que se adaptam a cada estudante em particular.

Na literatura existem diversos trabalhos que usam diferentes técnicas de Mineração de Dados no contexto educacional. Singh & Kumar (2012), por exemplo, utilizaram a técnica de mineração de dados chamada árvore de decisão para gerar conhecimento aos gestores da instituição para avaliar o desempenho de seus alunos. Dejaeger et. al. (2011), por outro lado, utilizaram a técnica de mineração de dados chamada clusterização de dados para identificar os principais fatores de satisfação dos alunos em duas instituições de ensino e consequentemente para a construção de modelos para apoiar os gestores no processo de tomada de decisão estratégica.

Diante os dados de avaliação dos alunos e com uso de Mineração de Dados, pode-se relacionar as questões socioeconômicas culturais dos estudantes com seu desempenho na prova, extraindo assim informações que possam ser úteis em planos estratégicos do desenvolvimento educacional e suas tendências de melhorias em seu setor.

O objetivo deste trabalho é a aplicação das etapas do KDD para o pre-processamento dos dados do ENEM e, consequentemente, o uso da técnica de Mineração conhecida por Associação de Dados. Este resulta em regras proposicionais que permitem a analise de causa e efeito e, dentro do contexto em que se insere o trabalho, o relacionamento do desempenho na prova com os fatores socioeconômicos.

Além da introdução que proporciona a motivação e contextualização do trabalho, compõe este trabalho as seguintes seções: Seção 3 uma breve introdução sobre

Mineração de Dados e detalhes do algoritmo de Associação Apriori; Seção 4 a metodologia empregada no trabalho, explicando detalhes da preparação da base de dados; Seção 5 os resultados e analises são apresentadas; e Seção 5, por fim, as conclusões.

3. Mineração de Dados

Por ser uma área multidisciplinar, envolvendo Banco de Dados, Estatística e Aprendizado de Máquina, as definições sobre Mineração de Dados podem variar de acordo com a área de aplicação. As definições abaixo foram retiradas das três áreas que são consideradas de grande expressão dentro da Mineração de Dados.

Cabena et al. (1998) apresentam uma definição sob uma perspectiva de banco de dados, sendo que a mesma é uma área de pesquisa interdisciplinar, atuando na interseção de técnicas de aprendizado de máquina, reconhecimento de padrões, estatísticas, banco de dados e visualização.

Hand et al. (2006) define sob uma perspectiva estatística, sendo um conjunto de técnicas uteis para analisar grandes bases de dados para a descoberta de padrões intrínsecos e apresentação dessa informação de forma compreensível aos tomadores de decisão.

Fayyad et al (1996) apresenta um viés de definição para um processo que é chamado de KDD, onde a Mineração de Dados consiste na principal etapa, entremeada a um pre-processamento dos dados originais e a um pós-processamento para analise dos padrões descobertos pela Mineração.

Na prática a Mineração de Dados é um termo usado amplamente para a descoberta de padrões, mas que porém, ela é composta por diferentes tarefas que podem atuar isoladamente ou em conjunto. Na Tabela 1 apresentamos as diferentes tarefas de Mineração de Dados e também as definições e exemplos de uso.

Neste trabalho adotaremos a tarefa de Associação de Dados. Na literatura temos dois algoritmos para esta tarefa, o A Priori e o FP-Grouth. Ambos mantem dois procedimentos que são a descoberta de itens conjuntos e a geração de regras de associação. A diferença principal é a eficiência na implementação dos algoritmos. O A Priori na etapa da verificação de itens conjuntos é feito por um processo iterativo para a combinação de itens. O FP-Grouth, por outro lado, tem uma estrutura de dados que permite o encontro de itens frequentes em uma única iteração. Neste trabalho adotaremos o A Priori e os detalhes desse algoritmo são apresentados na seção seguinte.

3.1 Associação de Dados Com o Algoritmo A Priori

Uma Regra de Associação consiste em uma expressão de implicação, representada como X → Y. Nessa relação, X e Y, antecedente e consequente, respectivamente, são conjuntos disjuntos de itens. A descoberta de um padrão de implicação pode existir pelo acaso, logo, é necessário medir a veracidade da Regra de Associação descoberta. Essa verificação, segundo Tan et. al. (2009), é realizada por meio dos seguintes elementos:

- Suporte: Determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados. É uma medida importante, pois é utilizada para eliminar regras sem interesse, bem como verificar a existência de uma regra que tenha baixo suporte por coincidência.

- Confiança: Determina a frequência na qual os itens em Y aparecem em transações que contenham X. Ou seja, mede a confiabilidade da inferência feita por uma regra. Caso se tenha uma determinada regra X → Y, o nível de confiança determina a probabilidade. Nesse caso, quanto maior a confiança, maior a probabilidade de que Y esteja presente em transações que contenham X. A confiança também fornece uma estimativa da probabilidade condicional de Y dado X.

Tabela 1: Tarefas de Mineração de Dados

Tarefa	Descrição	Exemplos
Classificação	Constrói-se um modelo com base a um conjunto de dados descritos por atributos e classes para que possa ser aplicado a dados não classificados. O objetivo é descobrir um relacionamento entre os atributos regulares (medições conhecidas) e um atributo especial (valor a ser previsto) de natureza categórica ou numérica discreta.	Classificar pedidos de crédito; Classificar operações fraudulentas; Classificar pacientes; Classificar estudantes.
Previsão (ou Regressão)	O mesmo principio da classificação de dados, com a diferença que o atributo especial tem natureza numérica contínua.	Prever o valor de vida de um equipamento; Prever o desempenho do aluno; Prever o índice de uma bolsa de valores; Prever a demanda de um consumidor para um novo produto.
Regras de Associação	Usada para determinar quais itens tendem a ser adquiridos juntamente em uma mesma transação.	Determinar produtos colocados juntos em um carrinho de supermercado; Determinar quais clientes compram dois produtos distintos; Determinar quais disciplinas o aluno tem desempenho semelhante.
Clusterização	Processo de partição de um conjunto de dados heterogêneos em grupos homogêneos.	Agrupar clientes por região do país; Agrupar clientes com comportamento de compra similar; Agrupar seções de usuários Web; Agrupar estudantes com desempenho semelhante.

3.1.1 O Algoritmo Apriori

O algoritmo Apriori foi o primeiro a ser utilizado para mineração de conjunto de itens ou simplesmente *itemsets* (do inglês) e regras de associação [Agrawal & Srikant, 1994]. Baseia-se no princípio de que qualquer subconjunto de itens frequentes deve ser um *itemset*. O algoritmo Apriori utiliza uma estratégia de busca em largura com um algoritmo de geração e teste. Em cada nível são gerados os *itemsets* possíveis, tendo em conta os mais frequentes gerados no nível anterior. Após serem gerados, a frequência desses *itemsets* é testada, percorrendo novamente a base de dados de transações [Facelli et al., 2011].

O algoritmo realiza busca por largura, e pode ser definido em três características: K-*itemsets*, suporte mínimo e confiança mínima. Define-se o valor mínimo do suporte para que um K-itemset seja considerado e a confiança limita a filtragem das associações descobertas pelo algoritmo. Para que o algoritmo Apriori

funcione corretamente, é necessário seguir todos os passos corretamente, como descrito na Tabela 2:

Tabela 2: Algoritmo A Priori

1. Dados associados transformados e/ou reorganizados, suporte mínimo e confiança mínima como dados de entrada.
2. Criar K-itemsets, considerando K=1.
3. Após análise dos dados, criar uma tabela com os K-itemsets de suporte maior que o definido como mínimo.
4. Criar um conjunto de candidatos a (K+1) itemsets a partir dos itemsets filtrados e utilizar as propriedades do algoritmo Apriori para eliminar os infrequentes.
5. Repetir os dois últimos passos até que o conjunto gerado seja vazio.
6. Listar as regras de associação e aplicar limite de confiança.

4. Metodologia

O ENEM é determinado por uma concepção construtivista, de maneira que suas provas têm sido elaboradas a partir de um forte foco na resolução de problemas, e não no mero exercício repetitivo de esquemas já bem aprendidos pelos estudantes [Macedo, 2005]. Além do foco na solução de problemas, há também um forte princípio de que a prova do ENEM não deve envolver significativamente a memorização ou a mera rapidez de raciocínio dos estudantes, mas valorizar a capacidade dos alunos em relacionar as informações dispostas pelo próprio item. Esse princípio enfatiza a capacidade do estudante em estabelecer conexões e lidar com questões que sejam verdadeiros desafios, o que é diferente do mero exercício, onde os itens já fazem parte do arsenal de exercícios prévios dos alunos, e já se sabe a resposta de antemão, de forma a não haver nenhum aspecto diferenciado que desafie a coordenar novas condições, projetar novas possibilidades, estratégias ou planos [Fini, 2005].

É uma prova única aplicada anualmente, e é composta por 180 (cento e oitenta) questões objetivas de múltipla escolha e uma redação, com sua referida proposta, e opcionalmente, caso o aluno queira preencher a fim de melhorar a qualidade do país, mais 200 (duzentas) questões socioeconômicas.

Para este trabalho utilizamos o banco de dados com os desempenhos da prova e do questionário socioeconômico do Enem de 2010. 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 4.33 Gigabytes.

Para acessar o banco de dados foi utilizado o software Oracle Express Edition 11g e o PL/SQL Developer, dois softwares que permitem a extração dos dados do arquivo que foi determinado para o foco da pesquisa, importando para um banco de dados apropriado para analise. Os dados que utilizamos para a mineração foram das capitais da região Sudeste: São Paulo, Rio De Janeiro, Belo Horizonte e Vitória.

Aproximadamente 4.200.000 (quatro milhões e duzentos mil) alunos de todo o Brasil estão inscritos e registrados no banco de dados do ENEM 2010. No questionário socioeconômico existem aproximadamente 200 (duzentas) questões a serem respondidas opcionalmente pelo inscrito, sendo assim, a base de dados possui o tamanho aproximado de 8.5 Gigabytes.

Na fase de pré-processamento, foram selecionados somente os dados das capitais da região Sudeste, resultando em 452.710 alunos. Continuando na mesma fase, foram eliminados todos os alunos da região que não compareceram nos dois dias da prova, resultando em aproximadamente 310.000 (trezentos e dez mil) alunos das quatro capitais.

As questões escolhidas para o foco da pesquisa foram as seguintes:

- Questão "1" – "Quantas pessoas moram com você?".
  - Alternativas: Uma a três "TRE", quatro a sete "SET", oito a dez "DEZ", mais de dez "MDEZ", moro sozinho "SOZ".
- Questão "2" – "Qual é o nível de escolaridade da sua mãe?".
  - Alternativas: analfabeto "ANA", 1º grau "1GR", 2º grau "2GR" e superior acima "SUP". Inicialmente, também se selecionou a questão que indagava a escolaridade do pai, porém o resultado se mostrou similar ao primeiro. Optou-se, então, por utilizar apenas a escolaridade relativa à mãe.
- Questão "3" – "Somando a sua renda com a renda das pessoas que moram com você, quanto é, aproximadamente, a renda familiar mensal?".
  - Alternativas: Até um salário mínimo "UM", de 1 a 3 salários mínimos "TRE", de 3 a 6 salários mínimos "SEI", de 6 a 9 salários mínimos "NOV", de 9 a 12 salários mínimos "DOZ", de 12 a 15 salários mínimos "QUI", mais de 15 salários mínimos "MQUI", Nenhuma renda "NER".
- Questão "4" – "Em que tipo de escola você cursou o Ensino Médio?"
  - Alternativas: Escola pública "PUB" e escola privada "PAR".

Estas questões foram escolhidas de forma a analisar os seguintes aspectos:

- A quantidade de pessoas que moram com o aluno possui algum tipo de interferência na nota do aluno no ENEM?
- O grau de escolaridade da mãe tem relação com o desempenho do aluno na prova objetiva?
- O valor da renda familiar mensal contribui para o desempenho do aluno?
- O tipo de escola em que o aluno estudou no Ensino Médio afeta o seu desempenho na prova?

Para efeitos de analise deste trabalho, a nota da prova objetiva foi classificada nos conceitos da Tabela 3.

Baseado nas respostas deste conjunto de questões selecionadas é possível analisar o desempenho do aluno às suas condições familiares e financeiras.

Após ter selecionado todos os critérios, foi desenvolvido uma limpeza e integração dos dados para preparação da aplicação da regra de associação, transformando as colunas principais em binominais, ou seja, 0 (zero) e 1 (um), sendo 1 (um) a opção respondida no questionário e a sua avaliação na prova, e 0 (zero) a alternativa que não se escolheu no mesmo questionário, juntamente com as avaliações

da prova em que o aluno não foi inserido, sendo obviamente três de quatro delas, conforme uma amostra apresentada na Tabela 4.

Tabela 3: Relação de transformação nota em conceito – prova objetiva

Nota da prova objetiva	Atribuição de conceito - prova objetiva
0.0 a 4.0	Insatisfatório - "INS"
4.01 a 6.0	Regular - "REG"
6.01 a 8.0	Bom - "BOM"
8.01 a 10.0	Excelente - "EXC"

Tabela 4 – Amostragem de alguns dos dados limpos e integrados

NU_INSCRICAO	UFSP	UFRJ	UFVI	UFBH	Q01TRE	Q01SET	Q01DEZ	Q01MDEZ	Q01SOZ	Q03ANA	Q031GR	Q032GR	NOTAREG
200000046162	0	0	0	1	1	0	0	0	0	0	0	0	1
200000045832	0	0	0	1	0	1	0	0	0	0	1	0	1
200000057142	0	0	0	1	1	0	0	0	0	0	0	1	1

Os critérios para identificação das colunas são as seguintes:

- NU_INSCRICAO: Identificação do candidato.
- UFXX: Sendo XX a sigla da cidade onde reside o candidato.
- QXXABCD: Sendo XX o número da questão escolhida para a pesquisa, de acordo com o questionário socioeconômico, e ABCD a alternativa da questão em forma abreviada, conforme classificado anteriormente.
- NOTAXXX: Sendo XXX a classificação da nota obtida pelo candidato.

5. Experimentos e Analises de Resultados

Os experimentos foram feitos com o uso da ferramenta RapidMiner 5.1 que utiliza código Java, open-source, que de acordo com MIERSWA (2006) fornece a implementação de algoritmos utilizados em problemas de aprendizagem de máquina e uma interface gráfica para o desenvolvimento rápido de projetos para a criação de modelos preditivos. É possível também definir processos de tratamento dos dados, inserindo para cada responsabilidade um operador como: entrada e saída, algoritmos de aprendizagem (supervisionado ou não), pré-processamento, validação e visualização.

A Tabela 5 apresenta as parametrizações de suporte e confiança mínimos utilizadas para o algoritmo A Priori e, também, o número de regras geradas em quatro simulações. Observa-se destes resultados que à medida que o suporte e a confiança decrescem, aumenta-se o número de regras.

Analisando as regras geradas em cada simulação, na Tabela 6 estão as de “S1”, as quais são interpretadas da seguinte maneira. Se o aluno estudou em escola pública então sua nota será avaliada como Regular na prova com 76% de confiança. Isto acontece em um total de 53% dos alunos. Ou seja, desse total de alunos das capitais da região Sudeste que estudaram em escola pública, 76% obtiveram uma nota regular na prova. Se a nota do aluno foi avaliada como regular na prova, então ele estudou em escola pública com 80% de confiança. Isto acontece em um total de 53% dos alunos. Ou seja, desse total de alunos que obtiveram uma nota regular na prova, 80% deles estudaram em escola pública.

Tabela 5 Simulações de Suporte e Confiança

Simulação	Suporte Mínimo	Confiança Mínima	Número de Regras Geradas
S1	50%	70%	2
S2	30%	70%	24
S3	25%	70%	42
S4	15%	70%	214

Tabela 6: Suporte Mínimo 50% e Confiança Mínima 70%

X	Y	Suporte	Confiança
PUB	REG	0,53	0,76
REG	PUB	0,53	0,80

Com a redução do suporte e a manutenção da confiança, houve aumento de regras e descoberta de relações de outros itens. Com o suporte mínimo de 30% e confiança mínima de 70% (simulação "S2" da Tabela 7), a primeira regra com menor confiança gerada é de que, se o estudante mora com quatro a sete pessoas então estuda em escola pública com 74% de confiança, ocorrendo com 31% dos alunos, como se pode observar na Tabela 7. A última regra com maior confiança é de que, se os pais tiveram escolaridade até o primeiro grau então o aluno estuda em escola pública com 89% de confiança, ocorrendo com 36% dos alunos.

Tabela 7: Suporte Mínimo 30% e Confiança Mínima 70%

X	Y	Suporte	Confiança
SET	PUB	0,31	0,74
1GR	PUB	0,36	0,89

Para a simulação "S3" da Tabela 8, a primeira regra com menor confiança gerada é de que, se os pais do estudante tiveram escolaridade até o primeiro grau então o aluno estudou em escola pública e sua nota foi regular na prova com 71% de confiança, ocorrendo com 28% dos alunos. A última regra com maior confiança é de que, se os pais têm escolaridade até o primeiro grau e o aluno estudou em escola pública então sua nota foi regular na prova com 90% de confiança, ocorrendo com 28% dos alunos.

Para a simulação "S4" da Tabela 9, podemos observar que há alguns atributos que levam ao aluno obter nota regular ou estudar em escola pública como a escolaridade dos pais e a renda da família. Podemos evidenciar na penúltima regra que, se o aluno tirou nota regular na prova, os pais tiveram escolaridade até o primeiro grau e a renda da família for de um a três salários mínimos então o aluno estudou em escola pública com 92% de confiança, ocorrendo em 19% dos alunos, como mostra a Tabela 9.

Sintetizando os resultados do algoritmo A Priori na Tabela 10, verificamos que, ao diminuir o suporte, há um crescimento no número de regras e, consequentemente, uma confiança maior nas regras geradas. Ou seja, quando diminuímos o valor do suporte mínimo, há um aumento nos casos encontrados na base de dados e, desses, utilizando a associação de dados, regras mais verdadeiras, com pelo menos 90% de confiança, são geradas na simulação correspondente. Nesta tabela, criamos faixas de confiança, em intervalos de 10% e a partir da confiança de 70%. A cada faixa, verificamos dentro das simulações a quantidade de regras geradas. Por exemplo, a Tabela 6, com resultados da simulação S1, temos duas regras, sendo uma com 76% de confiança e outra com 81%; portanto, uma se enquadra na faixa de 70% até 80% totalizando 50% - e a outra na faixa de 80% a 90% - totalizando 50%. Como resultado

de análise desta Tabela 9, o suporte representa a quantidade de atributos que ocorrem juntos. Como a proposta foi relacionar atributos de aspectos sociais com de desempenho na prova do ENEM, as intersecções com confiança grande aparecem com Suporte de 25% e de 15% da base de dados.

Tabela 8: Suporte Mínimo 25% e Confiança Mínima 70%

X	Y	Suporte	Confiança
1GR	PUB, REG	0,28	0,71
REG,1GR	PUB	0,28	0,90

Tabela 9: Suporte Mínimo 15% e Confiança Mínima 70%

X	Y	Suporte	Confiança
2GR	PUB	0,22	0,70
REG,TRE, UFSP	PUB	0,15	0,93

Tabela 10: Sintetização das Simulações.

	Suporte 50%	Suporte 30%	Suporte 25%	Suporte 15%
Confiança	S1	S2	S3	S4
70%≤min<80%	50%	50%	53,3%	44%
80%≤min<90%	50%	50%	40%	40%
90%≤min<100%	0%	0%	6,7%	16%

Com base nas regras de maior confiança geradas pelas duas últimas simulações, verificou-se que estudantes residentes na cidade de São Paulo com quatro a sete pessoas na moradia, de renda familiar de um a três salários mínimos, seus pais com a escolaridade até o primeiro grau e, no ENEM, sua nota obtida foi regular, então o aluno estuda em escola pública, reforçando que o nível da educação pública está ruim nas capitais da região Sudeste.

6. Conclusão

A partir dos resultados e do conhecimento extraído, a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas que moram com o estudante são atributos que diminuem o desempenho do aluno.

A educação no ensino público do Brasil ainda necessita ser melhorada, tanto politicamente quando pedagogicamente, e a classe social baixa é a mais afetada, influenciando diretamente no desempenho do estudante.

Em trabalhos futuros, pretende-se estudar a evolução do desempenho dos alunos na região Sudeste do Brasil e também nas outras regiões, com a base de dados das provas dos próximos anos do ENEM, para ser um eficiente indicador de educação.

7. Referências

Agrawal, R.; Srikant, R. (1994) “Fast algorithms for mining association rules”, In: *International Conference on Very Large Databases*, pages 487-499.

Cabena, P; Hadjinian, P; Stadler, R; Jaapverhees; Zanasi A. (1998), Discovering Data Mining: From Concept to Implementation, Prentice Hall, 1st edition.

Dejaeger, K., Goethals, F., Giangreco, A., Mola, L. and Baesens, B. (2011), "Gaining insight into student satisfaction using comprehensible data mining techniques", European Journal of Operational Research, Vol. 218(2), pages. 548-562.

Facelli, K, Lorena, A. C., Gama J. and Carvalho, A. P. D. L (2011), Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. Editora LTC.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Fini, M. E. (2005), Erros e acertos na elaboração de itens para a prova do ENEM. In: Ministério da Educação/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Exame Nacional do Ensino Médio (ENEM): fundamentação teórico-metodológica. Brasília: MEC/INEP. pages 101-106.

Hand, D; Mannila, H; Smyth, P (2001). Principles of Data Mining. MIT Press.

Macedo, L. de (2005). A situação-problema como avaliação e como aprendizagem. In: Ministério da Educação/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Exame Nacional do Ensino Médio (ENEM): fundamentação teórico-metodológica. Brasília: MEC/INEP. pages 29-36

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006, August). Yale: Rapid prototyping for complex data mining tasks. InProceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 935-940). ACM.

Paiva, R., Bittencourt I. I., Pacheco H., Silva A. P., Jacques P., Isotani S. "Mineração de dados e a gestão inteligente da aprendizagem: desafios e direcionamentos", XXXII Congresso da Sociedade Brasileira de Computação, 2012, Curitiba. Anais do DEsafIE! - I Workshop de Desafios da Computação Aplicada à Educação, 2012.

Romero, C., & 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.

Santos, M. C. D. (2013), Recursos de Tecnologia da Informação no Cenário Educacional: Princípios e Estratégias para Docentes Digitais. IX Simpósio de Excelência em Gestão e Tecnologia, 2012. Disponível em < http://www.aedb.br/seget/artigos12/19616322.pdf>. Acesso em 03 set. 2013.

Singh, S. and Kumar, V. (2012), "Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education", (IJCSN) International Journal of Computer Science and Network, Vol. 1(4), pages.121-126, ISSN: 2277-5420, 2012.

Tan, P.-N.; Steinbach, M. and Kumar, V. (2009). Introdução ao Data Mining – Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna Ltda.