Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio

Jéssica Laísa Dias da Silva, Isabel Dillmann Nunes

Curso de Sistemas de Informação – FACISA/CESED
Av. Argemiro de Figueiredo 1901, Itararé – 58.411-020 – Campina Grande-PB – Brasil
{jessicalaisajl, beldillnunes}@gmail.com

Abstract. *The search to find and understand the factors that can influence student learning is the concern of engineers and teachers. The Educational Data Mining uses techniques that offer an analysis of students, allowing identify several situations, such as dropouts or disapprovals. The aim of this article is to analyze a database of high school students by series, from the use classification technique using the J48 algorithm. This work allows us to understand the importance of classification technique of educational data mining, generates benefits and contributes to observation and classification of approved and disapproved students.*

Resumo. *A busca por encontrar e compreender os fatores que possam influenciar a aprendizagem do aluno é a preocupação de coordenadores e professores. A Mineração de Dados Educacionais utiliza técnicas que oferecem uma análise dos alunos, permitindo identificar várias situações, tais como desistências ou reprovações. O objetivo deste artigo é analisar uma base dados de alunos do Ensino Médio, por Série, a partir da utilização da técnica de Classificação usando o Algoritmo J48. Este trabalho permite perceber a importância da técnica de classificação da mineração de dados educacionais, gera benefícios e contribui na observação e classificação dos alunos aprovados e reprovados.*

1. Introdução

O Ministério da Educação relata que o ensino médio no Brasil é a etapa final da educação básica que deve compor a formação necessária para possibilitar, aos brasileiros enfrentar melhores condições de vida. Surge em controvérsia aos objetivos do Ministério da Educação, problemas como de evasão dos alunos, rendimentos escolares baixos e alunos que saem das instituições com dificuldades de aprendizados. Segundo pesquisa realizada pelo Ibope em 2011, 62% das pessoas com ensino médio não são plenamente alfabetizadas. A expectativa era que, aos 18 anos, e tendo frequentado a escola durante a infância e a adolescência, os jovens soubessem ler e entender textos longos, mas só 38% são capazes [Rodrigues 2014].

Diante destes problemas, pesquisadores buscam meios para identificar alunos em risco de reprovação em tempo para motivá-los e torná-los atuantes novamente. A Mineração de Dados Educacionais é uma das áreas que busca encontrar formas de melhorar a aprendizagem identificando os alunos em risco.

DOI: 10.5753/cbie.sbie.2015.1112

Segundo Baker *et* al. (2011) a Mineração de Dados Educacionais (do inglês *Educational Data Mining*, EDM) tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais.

Assim, visando analisar situações e características dos alunos de ensino médio, utiliza-se da técnica de classificação de Mineração de Dados Educacionais, e a utilização do algoritmo J48, aplicando-a a um conjunto de dados de uma instituição de ensino médio.

O algoritmo J48, segundo Martins et al. (2009), possibilita a criação de modelos de árvore de decisão. O modelo de árvore de decisão é construído pela análise dos dados de treino e o modelo utilizado para classificar dados ainda não classificados. O J48 gera árvores de decisão em que cada nó da árvore avalia a existência ou significância de cada atributo individual.

O presente trabalho tem como objetivo geral analisar uma base dados de alunos do Ensino Médio, por Série, a partir da utilização da técnica de Classificação usando o Algoritmo J48.

O artigo está estrutura da seguinte forma: o capítulo 2 mostra o conceito de Mineração de Dados Educacionais e uma revisão dos trabalhos relacionados, o capítulo 3 mostra a metodologia aplicada ao trabalho, os resultados são discutidos no capítulo 4 e por fim o capítulo 5 mostra as conclusões.

2. Mineração de Dados Educacionais

A Mineração de Dados Educacionais (do inglês, *Educational Data Mining,* que tem a sigla EDM) é conceituada como uma área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais, ou seja, esta área adapta métodos e algoritmos de Mineração de Dados existentes na literatura para explorar dados originados de ambientes educacionais [Baker et al. 2011].

A natureza destes dados é a mais diversa, exigindo adaptações e novas técnicas. Ao mesmo tempo, esta diversidade nos dados representa um grande potencial de implementação de recursos fundamentais para auxílio na melhoria da Educação. [Romero et al. 2008].

As técnicas são apresentadas conforme sua categorização nas subáreas de EDM, seguindo-se o que consta na taxonomia proposta por Baker et al. (2011). As técnicas de EDM são: Predição, Agrupamento, Mineração de Relações, Mineração de Regras de Associação, Mineração de Correlações, Mineração de Padrões Sequenciais, Mineração de Causas entre outras.

Romero e Ventura (2007) destacam que técnicas que se enquadram no grupo "Predição" estão sendo muito utilizadas. Assim nesse grupo enquadram-se as técnicas que têm por objetivo desenvolver modelos que possam inferir um aspecto particular dos dados (variável a ser prevista) por meio de alguma combinação de outros aspectos (variáveis preditoras).

Das técnicas citadas é utilizada neste trabalho a de predição, que trata-se de desenvolver modelos que deduzam aspectos específicos dos dados, em que ela visa prever o valor futuro de um determinado atributo.

2.1. Técnica de Predição com ênfase na Árvore de Decisão

Conforme Costa *et* al. (2012) a predição é um tipo de técnica que pode contribuir no desenvolvimento e uso de atividades instrucionais, devido ela conseguir estimar os benefícios educacionais antes mesmo da atividade ser aplicada aos alunos.

A Classificação é a técnica de predição a ser utilizada neste trabalho, visto que os estudantes são distribuídos em classes categóricas como por exemplo nota, frequência e perfil [Romero e Ventura 2007]. Romero et al. (2008b) ressaltam que existem diversos métodos classificação, dentre os quais: classificação estática, árvore de decisão, regras de classificação e redes neurais.

Entre os métodos de classificação optou-se por utilizar árvore de decisão pois é uma estrutura que pode ser utilizada para, por meio de uma regra, dividir sucessivamente uma grande coleção de registros em conjuntos menores. A cada divisão realizada, os dados são separados de acordo com características em comum até chegar a pontos indivisíveis, que representam as classes [Ossamu 2011].

Para compreender como é gerada uma árvore de decisão é necessário entender cada ponto que compõe sua estrutura. Segundo Coelho (2015), existem três tipos de nós: nó raiz que indica o inicio da árvore; ramos que dividem um determinado atributo e geram ramificações e nós folhas que contém as informações de classificação do algoritmo.

A mineração de modelos de classificação em bases de dados é um processo composto por duas fases: aprendizado e teste. Na fase de aprendizado, um algoritmo classificador é aplicado sobre um conjunto de dados de treinamento. Como resultado, obtém-se a construção do classificador propriamente dito. Deste modo cada observação do conjunto de treinamento é caracterizada por dois tipos de atributo: os atributos preditivos e os de classe, em que; e os atributos preditivos, os valores serão analisados para que seja descoberto o modo como eles se relacionam entre si e o de classe corresponde a qual a observação pertence [Martins et al 2009].

Vale ressaltar que uma árvore de decisão é estruturada também por um conjunto de regras de classificação. E que cada caminho da raiz até uma folha representa uma destas regras. Assim esta regra, deve ser definida de forma que, para cada observação da base de dados, haja um e apenas um caminho da raiz até a folha.

O algoritmo escolhido neste trabalho é J48, que baseia-se no algoritmo de árvores de decisão C4.5. Segundo Tavares et al.(2005) a forma de construção do J48 é a abordagem *top-down*, em que o atributo mais significativo, quando comparado a outros atributos do conjunto, é considerado raiz da árvore. Na sequência da construção, o próximo nó da árvore será o segundo atributo mais significativo, e, assim, sucessivamente, até gerar o nó folha, que representa o atributo alvo da instância.

2.2. Trabalhos Relacionados

As pesquisas em Mineração de Dados Educacionais vêm oferecendo contribuições significativas para a teoria e a prática da educação. No trabalho de Baker et al. (2011) são citados diversos exemplo do uso de métodos da EDM, para melhorar os modelos de conhecimento do estudante em vários diferentes domínios como ensino de língua estrangeira, geometria, química, física e outros. Um dos benefícios desse avanço foi a redução considerável do tempo gasto pelos alunos para desenvolver suas habilidades acadêmicas, principalmente em domínios como a matemática. Neste trabalho o autor traz também o exemplo de modelar emoções do estudante, permitindo verificar se um aluno está desmotivado ou confuso, objetivando a melhoria do *Design* Instrucional.

No Trabalho de Gottardo (2012) foi investigado como os dados armazenados por um AVA (Ambiente Virtual de Aprendizagem) poderiam ser transformados em informações potencialmente úteis para apoiar o acompanhamento de estudantes em cursos de ensino a distância (EAD). As informações geradas foram inferências envolvendo estimativas de desempenho acadêmico futuro de estudantes. Os resultados obtidos com a aplicação de técnicas de mineração de dados sobre o conjunto de atributos selecionados demonstram que é possível obter inferências relativas ao desempenho dos estudantes com taxas de acurácia global variando entre 72% e 80%.

Rigo et al. (2012) realizou uma análise de melhorias possíveis na aplicação de Mineração de Dados Educacionais, em que seus resultados pudessem apoiar efetivamente processos de detecção de comportamentos ligados à evasão escolar. Assim tratou da importância de realizar um mapeamento que viabiliza ver os problemas da escola analisando as possíveis melhoras. Constataram que dada a diversidade dos dados envolvidos, existiam possibilidades para utilização de conjuntos combinados de técnicas de Mineração de Dados Educacionais. Esta exploração de algoritmos, técnicas e mecanismos foi destacada com um dos pontos cruciais para que sejam alcançados os resultados.

Kampff (2008) identifica por intermédio da Mineração de Dados o comportamento e as características de alunos propensos à reprovação e à evasão por meio de uma arquitetura para um sistema de acompanhamento das iterações em um Ambiente Virtual de Aprendizagem. O professor recebe um alerta através do sistema consentindo estabelecer comunicação personalizada e contextualizada com os alunos. Ao fim, foi possível evidenciar que as intervenções realizadas pelo professor, a partir dos alertas, colaboraram para o progresso dos índices de aprovação e para redução dos índices de evasão dos alunos na disciplina. Contudo, o sistema de acompanhamento desenvolvido utiliza como parâmetro de entrada um arquivo cvs, exportado do banco de dados, o que pode gerar ruídos e a não integração entre o banco de dados do Ambiente Virtual de Aprendizagem e o sistema de acompanhamento.

Enfim cada vez mais diversas soluções vêm sendo desenvolvidas com o estudo de métodos da Mineração de Dados Educacionais na literatura, com aplicações para: apoiar os tutores e responsáveis educacionais, ao fornecer estatísticas de uso do sistema e feedback sobre a iteração dos alunos.

3. Metodologia

A proposta deste trabalho é aplicar o algoritmo de classificação J48 (utilizando a ferramenta Weka[1]) aos dados dos alunos de ensino médio dos anos de 2011 a 2014 de uma escola particular de ensino médio. E assim realizar uma análise de padrões e características dos alunos que estiverem em risco de reprovação.

A Figura 1 ilustra os passos da posposta deste trabalho:

- Os professores e coordenadores alimentam o sistema acadêmico da instituição com os dados (frequência, nota, série, ano, perfil do aluno, ...).
- Extrair os dados: neste ponto é filtrado apenas os dados que serão usados para aplicação da técnica de mineração de dados educacionais.
- Weka: neste ponto aplicando a técnica de árvore de decisão sob os dados
- Relatório Gerado pelo Weka: mostra o resultado obtido após aplicado o algoritmo na ferramenta.
- Visualização das informações- neste ponto é mostrado os resultados obtidos pelo classificador como também árvore de decisão gerada para assim permitir a análise da base de dado, observando os padrões e características dos alunos.

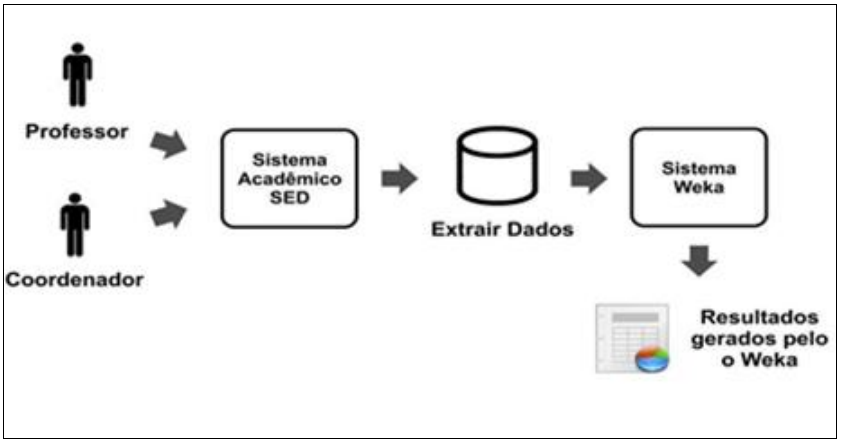

Figura 1 - Esquema do trabalho proposto

A base de dados obtida para realização deste trabalho foi de uma escola particular de Ensino Fundamental e Médio. Os dados referem-se aos alunos dos anos de 2011 a 2014 da 1ª, 2ª e 3ª séries. A quantidade de dados selecionada de cada nível pode ser visualizada na Tabela 1.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Tabela 1 – Tabela de quantidade de dados

Níveis	Total
1ª série	1649
2ª série	1289
3ª série	1087

Foram considerados os seguintes atributos de cada aluno: Nível (1ª, 2ª ou 3ª série); Cidade de origem (Local ou Vizinha); Bolsista (Sim ou Não); Desistentes (Sim ou Não); Sexo (F ou M); Idade (<=17 e >17) e Ano (2011 a 2014).

Com o intuito de observar o desempenho do algoritmo J48 são utilizados o percentual de Instâncias Corretamente Classificadas (ICC) e de Instâncias Erroneamente Classificadas (IEC).

4. Resultados

A aplicação do algoritmo de classificação J48 para a 1ª Série do Ensino Médio obteve ICC de 77.9867 % e IEC de 22.0133 %. A árvore de decisão resultante pode ser visualizada na Figura 2.

```
J48 pruned tree
------------------

DESISTENTE = NAO_DESISTENTE
|   ANO = 2011: APV (244.0/84.0)
|   ANO = 2012: APV (239.0/46.0)
|   ANO = 2013
|   |   IDADE = <=17: APV (463.0/99.0)
|   |   IDADE = >17
|   |   |   SEXO = F: APV (3.0)
|   |   |   SEXO = M: REPV (7.0/2.0)
|   ANO = 2014
|   |   BOLSISTA = NAO_BOLSISTA: REPV (333.0/125.0)
|   |   BOLSISTA = SIM_BOLSISTA: APV (5.0)
DESISTENTE = SIM_DESISTENTE: REPV (355.0)
```

Figura 2 - Árvore de decisão 1ª Série

A partir da árvore de decisão gerada, pode-se inferir que:

- No ano de 2011, foram classificados corretamente 244 registros de aprovados e 84 foram classificados incorretamente.
- No ano de 2012 foram classificados corretamente 239 registros de aprovados e 46 classificados incorretamente.

- No ano de 2011 e 2012 não tiveram nenhuma ramificação, apenas apresentou os registros dos alunos aprovados .
- No ano de 2013, teve um diferencial, pois é apresentado duas ramificações, uma para idade e outra para sexo. Observa-se ainda, que houve maior quantidade de aprovados os quais estavam na faixa etária menor igual a 17. Percebe-se que os maiores de 17 anos, do sexo masculino, tiveram maior número de reprovação.

Para a 2ª série foi visto que a técnica de estratificação obteve ICC de 71.6059 % e o IEC de 28.3941 %. A Figura 3 mostra a árvore de decisão obtida.

```
DESISTENTE = NAO_DESISTENTE
|   IDADE = <=17
|   |   ANO = 2011: APV (284.0/91.0)
|   |   ANO = 2012: APV (235.0/86.0)
|   |   ANO = 2013: APV (225.0/51.0)
|   |   ANO = 2014
|   |   |   SEXO = F
|   |   |   |   CIDADE = V: REPV (41.0/16.0)
|   |   |   |   CIDADE = CG
|   |   |   |   |   BOLSISTA = NAO_BOLSISTA: APV (113.0/45.0)
|   |   |   |   |   BOLSISTA = SIM_BOLSISTA: REPV (2.0)
|   |   |   SEXO = M
|   |   |   |   BOLSISTA = NAO_BOLSISTA
|   |   |   |   |   CIDADE = V: APV (7.0/1.0)
|   |   |   |   |   CIDADE = CG: REPV (101.0/36.0)
|   |   |   |   BOLSISTA = SIM_BOLSISTA: APV (4.0)
|   IDADE = >17
|   |   ANO = 2011
|   |   |   SEXO = F: REPV (10.0/4.0)
|   |   |   SEXO = M: APV (13.0/4.0)
|   |   ANO = 2012: REPV (27.0/7.0)
|   |   ANO = 2013: REPV (14.0/4.0)
|   |   ANO = 2014
|   |   |   BOLSISTA = NAO_BOLSISTA: REPV (26.0/6.0)
|   |   |   BOLSISTA = SIM_BOLSISTA: APV (3.0)
DESISTENTE = SIM_DESISTENTE: REPV (184.0)
```

Figura 3-Árvore de decisão 2ª Série

Analisando a árvore de decisão gerada da 2ª Série destaca-se:

- Árvore se divide na idade tendo duas ramificações uma para os alunos menores iguais a 17, e outra para faixa etária maior que 17.
- Na primeira faixa etária merece destaque o ano de 2011 que teve mais aprovação. Em seguida teve outra ramificação por sexo em que, o sexo masculino de Campina Grande teve mais número de reprovados comparado as cidades vizinhas.
- Em geral o sexo masculino teve mais reprovados que o feminino.
- número de desistente também foi pequeno apenas de 184 comparado ao número de alunos analisados.
- Na faixa etária maior que 17 é destacado, o número de reprovados.

- Outro ponto que merece evidência é que os números de desistentes são inferiores, comparado à quantidade de alunos (335).

Por fim para a 3ª Série do Ensino Médio, foi visto que a técnica de estratificação que obteve ICC de 77.3689 % e o IEC de 22.6311%. A Figura 4 mostra a árvore de decisão correspondente.

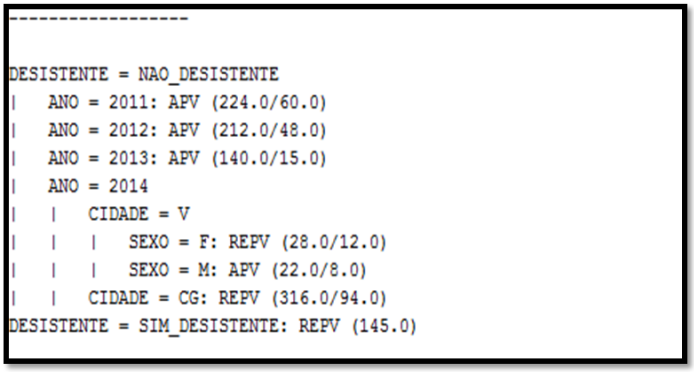

Figura 4- Árvore de decisão 3ª série

Análise feita na árvore de decisão da 3ª série:

- Dentre os anos 2011,2012 e 2013 o que obteve maior quantidade registros de aprovados foi o de 2011.
- Observa-se que os números de reprovados das cidades vizinhas é menor comparado com a cidade de Campina Grande.
- É percebido também que o número de desistentes é pequeno comparando a quantidade de registros de aprovados.

Assim, a partir da análise das árvores de decisãoé possível observar que na 1ª série tem uma maior quantidade de alunos em relação as demais, porém a 3ª série apresentou maior número de aprovados proporcionalmente.

Observaram-se algumas semelhanças para séries:

- Os alunos de cidades vizinhas são menos reprovados que os de Campina Grande;
- Os alunos bolsistas são todos aprovados;
- Número de desistentes é pequeno levando em consideração o total de aluno por série.

7. Conclusão

Neste trabalho foi ressaltado o quanto a mineração de dados educacionais é importante e como sua utilização pode contribuir para análises e aprendizagens sobre um conjunto de dados.

A técnica de Mineração de Dados Educacionais utilizada foi a de classificação, a qual permitiu compreender as informações armazenadas dos estudantes de ensino médio. O algoritmo J48 foi escolhido para classificar os dados. O J48 permite gerar uma árvore de decisão, estrutura essa que mostra os principais atributos que influenciam na classificação de estudantes aprovados e reprovados.

Algo também que merece destaque é a possibilidade de utilizar a técnica de classificação na ferramenta Weka, que possibilita fazer uma série de testes, gerando relatórios, que contribuem para encontrar a técnica mais adequada ao problema analisado. No caso deste trabalho procurou-se realizar análises comparativas com as técnicas de estratificação disponíveis: use training e set, supplied test set, cross-validation e percentage Split com uso de filtro e sem uso. Ao aplicar o filtro resample, em cada experimento, contribui com a qualidade e desempenho do classificador de cada série, pois ele realiza balanceamento das classes não permitido que o classificação fique tendenciosa a classificar só os alunos aprovados por exemplo.

Por meio da técnica utilizada da mineração de dados analisou-se a situação com a base dos alunos do ensino médio e verificou-se que a quantidade de alunos que são desistentes e bolsista é pequena. O número de alunos aprovados é maior que reprovados e os alunos de cidades vizinhas são menos reprovados que os de Campina Grande. Percebeu-se também que o número de reprovados na 3ª série é menor que nas demais.

Para trabalhos futuros, pretende-se buscar mais atributos para esta base, tais como: renda familiar, frequência, moradia, atividades extras e notas. Podendo assim fazer outras análises. Porque foi visto que quanto mais dados são disponibilizados, melhor desempenho tem o classificador e mais informações podem ser extraídas e consequentemente mais testes poderão ser executados. E também utilizar outros algoritmos disponíveis para técnica de classificação realizando testes comparativos do desempenho de cada um, verificando as vantagens e desvantagens deles sobre a base. Outra pretensão é gerar regras de classificação para solucionar problemas de rendimento escolar.

Referências

Baker, R. S. J., Isotani, S., Carvalho, A. de**.** (2011) "Mineração de dados educacionais: Oportunidades para o Brasil" Revista Brasileira de Informática na Educação, v. 12, n. 2, p. 3 – 13.

Coelho R. A. (2015). "Árvores de Decisão". http://www.ime.usp.br/~slago/sia-ad.pdf. Abril.

Costa E., Baker R. S. J., Amorim L., Magalhães J. e Marinho T. (2012) "Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações." Anais Jornada de Atualização em Informática na Educação.

Gottardo, E. (2012) "Estimativa de Desempenho Acadêmico de Estudantes em um Ava Utilizando Técnicas De Mineração De Dados". Dissertação de Mestrado na Universidade Tecnológica Federal Do Paraná. Programa De Pós-Graduação em Computação Aplicada. Curitiba.

Kampff, A. J. C. (2009) "Mineração de Dados Educacionais para a Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente." 186p,

Tese (Doutorado em Informática na Educação) – Programa de Pós Graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Martins, A. C.; Marques, J. M. e Costa, P. D. C. (2009) "Estudo Comparativo de três Algoritmos de Machine Learning na Classificação de Dados Electrocardiográficos". http://www.dcc.fc.up.pt/~ines/aulas/0910/MIM/trabs_ano_anterior/noname.pdf.

Ossamu E. H. (2011) "Técnica de Árvore de Decisão em Mineração de Dados" http://www.fatecsp.br/dti/tcc/tcc0003.pdf.

Rigo, S. J., Cazella, S. C. e Cambruzzi, W. (2012) "Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades." Anais do Workshop de Desafios da Computação Aplicada à Educação.

Rodrigues, C. (2011) "Alunos terminam ensino médio sem aprender". Disponível em: http://ultimosegundo.ig.com.br/educacao/alunos+terminam+ensino+medio+sem+aprender/n1238097714540.html.

Romero, C. e Ventura, S. (2007) "Educational Data mining: A Survey from 1995 to 2005" Expert Systems with Applications, v.33, p.125-146, 2007.

Romero, C., Ventura S. e Garcia, E. (2008) "Data mining in course management systems: Moodle case study and tutorial". Computers & Education, n. 51, p.368-384.

Romero, C., Ventura, S., Espejo, G. P. e Hervás, C. (2008b) "Data mining Algorithms to Classify Students". In Proceedings of the 1st International Conference on Educational Data mining, p.8-17.

Tavares, C., Bozza, D. e Kono, F. (2005) "Descoberta de Conhecimento aplicado a Dados Eleitorais". Revista Gestão & Conhecimento - *On-line*. v. 5, n. 1. Páginas 54 - 94.