5CONTECSI June 04 - 06, 2008 São Paulo, Brazil

5th International Conference on Information Systems and Technology Management
5º Congresso Internacional de Gestão da Tecnologia e Sistema de Informação
De 04 a 06 de Junho de 2008 - São Paulo - Brasil

**PS-991**

# DATA MINING AS A DECISION TOOL: CASE STUDY OF A UNIVERSITY

Bruno de Abreu Machado (Universidade Fumec - Face, Belo Horizonte, Brasil)
brunodeabreu@gmail.com
George Leal Jamil (Universidade Fumec - Face, Belo Horizonte, Brasil) gljamil@terra.com.br

Information needs are more perceptible in modern enterprises, as decision taking processes demand precise descriptions of businesses environments. This information usually is spread over multiple databases and there would be situations where it must be gathered or related for complex analysis. Among tools and computational techniques applied for information availability, one of the most used is data warehouse, which can be used as database or information repository for managerial purposes, serving for data mining processes which are designed to analyze and relate data aiming more accurate decisions. Data mining algorithms implement classification and analysis over data warehouses and generate hypothesis from analyzed sources. These hypotheses can indicate business trends or knowledge hidden in standards. This study, conducted on a University, showed that would be possible to apply mining techniques to classify admission candidate data in order to evaluate business alternatives, as regions for a communication campaign.
Keywords: Decision taking processes, data mining, data warehouse, database, knowledge discovery.

# MINERAÇÃO DE DADOS COMO FERRAMENTA DE TOMADA DE DECISÃO: ESTUDO DE CASO DE UMA UNIVERSIDADE

A necessidade por informação é cada vez mais visível nas empresas, o processo de tomada de decisão consome e demanda informações precisas do ambiente empresarial como um todo. Essas informações normalmente estão em diversas bases de dados e pode haver a necessidade de serem analisadas em conjunto e relacionadas. Dentre as ferramentas e técnicas computacionais auxiliam com diversas formas de disponibilizar essas informações, uma das formas mais utilizadas é o data warehouse que se torna um repositório de informações, podendo servir aos sistemas e processos de mineração de dados para analisar e relacionar os dados com objetivo de propiciar decisões de maior precisão. Os algoritmos de data mining implementam tarefas de classificação e análise sobre esta base e geram algumas hipóteses com base nos dados analisados. Essas hipóteses apontam alguma tendência ou relacionamento não facilmente perceptível, possibilitando diversas analises e descoberta de conhecimento. Esse estudo, conduzido em uma Universidade, mostrou que seria possível utilizar as técnicas de mineração para classificar seus candidatos num processo de vestibular e obter hipóteses de negócios, como quais são as regiões propicias a se investir em campanhas publicitárias.
Palavras - Chave: Tomada de decisão, mineração de dados, data warehouse, base de dados, descoberta de conhecimento.

# 1. INTRODUÇÃO

Para se destacar no mercado competitivo as empresas querem, atender melhor seu cliente, buscar novas tendências de mercado, aproveitar novas oportunidades e sobre tudo entender seu próprio negócio. O departamento de TI[1] pode auxiliar as empresas oferecendo ferramentas e técnicas de análise de dados para descobrir conhecimento não facilmente perceptível.

Dentre essas técnicas destaca-se o *data mining* ou mineração de dados, técnica que auxilia as empresas a descobrir conhecimento em bases de dados, hoje as ferramentas de *data mining* possuem algoritmos e técnicas que facilitam e organizam as informações facilitando sua compreensão e auxiliando a tomada de decisão.

Estas técnicas direcionam, auxiliam a traçar um plano de descoberta de conhecimento que inclui diversas etapas de pré-processamento, mineração de dados e pós-processamento que é a etapa onde se disponibiliza as informações para tomada de decisão.

Choo (2003) destaca que durante a tomada de decisão, as principais atividade são o processamento e análise da informação a partir das alternativas disponíveis, cujas vantagens e desvantagens são pesadas. O processo de *data mining* trata justamente do processamento das informações.

O volume de dados em uma empresa cresce rapidamente, com a ajuda de sistemas de gestão como ERP[2], CRM[3], ou outros feitos para atender atividades rotineiras, a grande quantidade de dados dificulta a interpretação e análise em um mesmo contexto, informações preciosas podem não ser percebidas, com isso, a etapa de processamento e análise se torna tão difícil, temos então ferramentas de *data mining* para auxiliar neste trabalho. Segundo Passos e Goldschmidt (2005, p.252), “a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas”.

Com essas ferramentas e técnicas o departamento de TI assume um papel estratégico dentro da organização, oferecendo informações valiosas e muitas vezes decisivas para tomada de decisão, crescendo a responsabilidade dos gerentes e analistas. Um dos grandes desafios de atender essa nova demanda é transformar grandes quantidades de dados em informações relevantes e extrair dessas informações o conhecimento necessário para resolver algum problema ou atender alguma expectativa.

As empresas possuem grandes quantidades de dados armazenados em diversos repositórios e, estes dados não são explorados de tal forma a oferecer informações relevantes onde podem ser tomadas decisões importantes, estas decisões devem estar aliadas e focadas no objetivo estratégico da empresa, bem como sua missão e seu plano estratégico previamente traçado (JAMIL,

[1] TI : Tecnologia da Informação

[2] ERP : *Enterprise Resource Planning*, sistema de gestão que permite as empresas automatizar e integrar processos operacionais

[3] CRM : *Customer Relationship Management*, sistema de gestão empresarial que onde seu foco é o cliente

2007), a questão que trataremos é, como o *data mining*, com suas técnicas e ferramentas de mineração de dados, pode auxiliar as empresas na tomada de decisão?

O setor de TI e seus profissionais podem contribuir e auxiliar em diversos processos estratégicos na empresa, deixa sua postura puramente técnica e assume a postura de Gerencia do Conhecimento da empresa, sendo decisivo para o sucesso de uma decisão, informações precisas e rápidas são imprescindíveis.

Técnicas de *data mining* podem ser aplicadas para gerar diversos ganhos como, traçar o perfil de um cliente que tem um bom relacionamento no mercado, verificar possíveis fraudes, análise de dados de vendas e ainda ganhos sociais como, identificar uma fatia da população que precise de maior assistência em um determinado ramo social.

Com isso, os dados se tornam um patrimônio importantíssimo para a empresa e explorá-los de forma sábia pode ser um fator decisivo, para o crescimento da empresa, sustentação de uma posição no mercado ou até mesmo decisões importantes de investimentos.

O objetivo é mostrar como as empresas podem aproveitar e utilizar as valiosas informações contidas em suas bases de dados, facilitando ou orientando a toma de decisão.

# 2. DESCOBRIMENTO DE CONHECIMENTO EM BASE DE DADOS

O processo de descoberta do conhecimento em base de dados é caracterizado como um processo composto por várias etapas operacionais, dentre elas o pré-processamento, mineração e o pós-processamento, entre essas fases também se destacam uma série de etapas.

Para Rezende (2006) o processo de extração do conhecimento em bases de dados tem como objetivo encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório.

O processo de descoberta do conhecimento se inicia no domínio do problema, saber qual é a situação problema que se deseja resolver ou analisar. Para Barbieri (2001) é importante que para se iniciar a descoberta de conhecimento algumas questões estejam resolvidas como o entendimento do negócio e suas metas, saber exatamente o que se necessita, entender o problema em riqueza de detalhes, deter de técnicas e ferramentas adequadas e possuir bons usuários para o projeto.

De acordo com Schenatz (2005) o processo de KDD[4] é interdisciplinar e envolve áreas relativas a aprendizado de máquinas, conhecimento de padrões, banco de dados, estatística e matemática, aquisição de conhecimento para sistemas especialistas e visualização de dados.

[4] KDD – *Knowledge Discovery in Database* : Descoberta de conhecimento em base de dados

O processo utiliza de técnicas, algoritmos e métodos das diversas áreas, para extrair conhecimento útil das grandes quantidades de dados, esse processo pode ser visto na figura a seguir.


Figura 01: Processo de KDD
Fonte: SCHENATZ, 2005, p.35

Com essa característica o processo de KDD se torna complicado por exigir do profissional uma gama muito grande de conhecimento, com isso, é importante que as diversas áreas da empresa estejam envolvidas no projeto de descoberta do conhecimento, para assim dividir as responsabilidades e aperfeiçoar o processo.

Segundo Goldschmidt e Passos (2005) para uma melhor compreensão do processo de KDD é necessária uma apresentação dos principais elementos envolvidos em aplicações nesta área, esse elementos são: o problema, os recursos disponíveis para solução do problema e os resultados obtidos.

Deve-se então para aplicar o KDD em uma organização primeiramente identificar e entender o problema, organizar e obter recursos para aplicação do processo e saber analisar os dados apresentados.

O processo de descoberta do conhecimento em bases de dados começa com o entendimento do domínio da aplicação e a relevância do conhecimento em relação as metas a serem atingidas. Em seguida é feita a seleção dos conjuntos de dados a serem utilizados durante o processo de KDD, isto é, um agrupamento organizado de dados, que será o alvo da prospecção. A etapa de limpeza dos dados (*data cleaning*) vem a seguir, por meio de um pré-processamento dos dados visando adequá-los aos algoritmos. (SCHENATZ, 2005 p.36)


Figura 02: Etapas do processo de KDD
Fonte: SCHENATZ, 2005, p.36

Na (Figura 02) temos uma visão geral das etapas do processo de descoberta do conhecimento e percebe-se que para o sucesso de uma mineração de dados temos uma série de etapas que devem ser executadas para que os algoritmos e as análises sejam executadas e nos ofereçam informações úteis.

Na etapa de seleção dos dados temos a tarefa de identificar quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD. (GOLDSCHMIDT; PASSOS, 2005, p.26)

Em geral os dados se encontram em base de dados transacionais que estão sendo constantemente utilizados, sofrendo atualizações e consultas, uma alternativa é que seja feita uma cópia dos dados a serem analisados e importados para uma tabela, outra possibilidade em caso de existência de *data warehouse* ou *data mart* que esse armazém de dados seja utilizado.

Segundo Goldschmidt e Passos (2005) a maioria dos métodos de mineração de dados pressupõe que os dados estejam organizados em uma única e possivelmente muito grande estrutura tabular dimensional.

Percebe-se então que o processo de mineração pode ocorrer independente da estrutura dos dados, onde pode estar em tabelas, *data marts* ou *data warehouse*, levando em consideração somente a sua organização e modelagem. Uma das melhores formas de selecionar os dados é a migração de todo o conteúdo, atributos e registros de uma tabela em uma base de dados relacional e a outra é a seleção detalhada, orientada dos campos e registros que irão compor a estrutura a ser analisada.

Considera-se que a etapa de pré-processamento compreende as funções relacionadas à captação, à organização, ao tratamento e à preparação dos dados para a etapa da Mineração dos Dados.

Goldschmidt e Passos (2005) destacam que esta etapa possui fundamental relevância no processo de descoberta de conhecimento. Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração a serem utilizados.

Analisando os conceitos pode-se considerar que a seleção dos dados pode fazer parte da etapa de pré-processamento, onde ocorre a preparação dos dados que serão processados.

Com relação a fase de mineração de dados Schenatz (2005, p.36) aborda que:

Na fase de mineração de dados ou *Data Mining* especificamente, que começa pela escolha do algoritmo a ser aplicado, essa escolha depende fundamentalmente do objetivo do processo de KDD: classificação, segmentação, agrupamento por afinidades, estimativas, árvores de decisão, etc. De modo geral, na fase de *Data Mining*, ferramentas especializadas procuram padrões nos dados. Essa busca por ser efetuada automaticamente pelos sistemas ou interativamente com um analista responsável pela geração da hipóteses. Diversas ferramentas distintas, como redes neurais indução de árvore de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente como em combinação, podem ser então aplicadas ao problema.

# 3. DATA MINING

O avanço da tecnologia tem nos proporcionado armazenar uma grande quantidade de dados, analisar esses dados é uma tarefa de extrema dificuldade. As empresas necessitam desses dados e a cada vez mais demanda uma visão global, macro desses dados.

"*Data mining* ou mineração de dados é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados, usando-as para efetuar decisões cruciais. O *data mining* vai muito além da simples consulta a um banco de dados, no sentido de que permite aos usuários explorar e inferir informação útil a partir dos dados, descobrindo relacionamentos escondidos no banco de dados. Pode ser considerada uma forma de descoberta de conhecimento em banco de dados". (GROTH, 1997, apud SCHENATZ, 2005, p.33)

O *data mining* consiste em buscar informações valiosas das bases de dados das empresas, visando auxiliar em um processo de tomada de decisão, para alguns autores o *data mining* é uma forma de KDD, já outros como com Goldschmidt e Passos (2005) consideram que a mineração de dados uma etapa do processo de descoberta do conhecimento em base de dados, onde ocorre a busca efetiva de conhecimento.

Com a utilização do *data mining* as empresas podem modificar sua forma de atuação ou revolucionar suas atividades como destacado por Bispo (1999) o setor de marketing também esta se revolucionando com o uso de *Data Mining*. Em vez de realizar imensas e caras campanhas de âmbito geral, essas organizações descobriram que, dividindo o público-alvo em categorias, é possível realizar campanhas mais direcionadas, mais baratas e com um retorno muito maior. Para dividir o público-alvo em categorias, é necessário conhecer esse público, e o *data mining* tem sido imprescindível nesse sentido.

Outros setores das empresas podem utilizar dessa tecnologia para melhorar seus processos de tomada de decisão, como o setor de compras, onde o *data mining* poderia auxiliar na escolha e seleção de potenciais fornecedores ou pontos ideais de ressuprimento de material, em um departamento de vendas a mineração de dados pode identificar um potencial comprador para um determinado produto e principalmente nas áreas gerenciais onde a mineração oferece informações importantes para tomada de decisão.

# 4. TÉCNICAS DATA MINING

O *data mining* pode executar uma série limitada de tarefas, dependendo das circunstâncias. Autores como Goldschmidt, Passo, Schenatz, Tang, Maclennan, Miranda, Reis e Arbex, falam sobre essas tarefas ou funções que envolvem processos de KDD e mineração de dados. Os algoritmos de dados visam satisfazer ou são classificados dentro de algumas dessas tarefas.

## Classificação

Uma das tarefas mais populares de KDD pode ser compreendida como a busca de uma função que permita associar corretamente cada registro **Xi** de um banco de dados a um único rótulo categórico, **Yi,** denominado classe. (GOLDSCHMIDT; PASSOS, 2005, p.66)

Para Schenatz (2005) consiste no mapeamento ou pré-classificação de um conjunto pré-definido de classes, normalmente algoritmos de classificação incluem árvores de decisão ou redes neurais.

No exemplo seguinte, a classe clientes e os dados (idade, se possui carro ou não, se realizou compra ou não) são descritos da seguinte forma:

SE *cliente* >= 40anos
ENTÃO *cliente compra*
SENÃO *cliente não compra*
SE *cliente* >= 30 e < 40 anos e não possui carro
ENTÃO *cliente compra*
SENÃO *cliente não compra*

Fonte: Machado,B.A (2007)

De acordo com Tang e Maclennan (2005) a classificação é uma das mais populares tarefas e se refere a associação de casos em categorias baseadas em atributos previsíveis, a tarefa requer encontrar um modelo que descreve a classe de atributos e o funcionamento da entrada de atributos.

## Regressão

De acordo com Goldschmidt e Passos (2005) a tarefa de regressão compreende fundamentalmente a busca por funções lineares ou não, que mapeiem os registros de um banco de dados em valores reais. Tarefa similar a de classificação sendo restrita a atributos numéricos.

A tarefa de regressão é semelhante a da classificação, contudo sua especialidade é analisar uma série continua de atributos previsíveis, “é uma tarefa amplamente estudada por estatísticos, técnicas de regressão podem resolver muitos problemas empresariais”, conforme Tang e Maclennan (2005, p.31).

Normalmente algoritmos de estatística e redes neurais oferecem suporte para implementar a regressão. (MICHIE et all, 1994, apud GOLDSCHMIDT; PASSOS, 2005, p.73)

## Previsão

De acordo com Schenatz (2005) os registros são classificados de acordo com alguma atitude futura prevista. Em um trabalho de previsão, o único modo de configurar a precisão é esperar para ver.

A tarefa de precisão pode nos ajudar a prever uma série de situações oferecendo vantagem competitiva e um valioso suporte a tomada de decisão, por exemplo, qual será a previsão de faturamento da empresa no próximo mês? Essa informação pode ser crucial para uma decisão de investimento.

A análise de uma série temporal é o processo de identificação das características, dos padrões e das propriedades importantes da série, utilizados para descrever em termos gerais o seu fenômeno gerador. Dentre os diversos objetivos de análises de séries temporais, o maior deles é a geração de modelos voltados à previsão de valores futuros. (GOLDSCHMIDT; PASSOS, 2005, p. 78)

Pretende-se com a previsão analisar um conjunto de dados em função do tempo, analisando um conjunto de dados em um determinado período.

## Descoberta de associações

Segundo Goldschmidt e Passos (2005) essa tarefa consiste em encontrar conjunto de itens que ocorram simultaneamente e de forma freqüente em um banco de dados.

Considerar-se uma tabela de vendas de uma determinada organização, aplicas-se o conceito de descoberta de associações para identificar os produtos mais vendidos de forma conjunta. Associar uma informação com a outra mesmo que essa associação não pareça muito lógica.

Goldschmidt e Passos (2005) citam o algoritmo Apriori dentre os exemplos de descoberta de associações.

## Analise de seqüências

A seqüência é uma extensão da tarefa de descoberta de associações, ela considera espaço temporal entre as transações registradas no banco de dados, ela considera uma seqüência de eventos ocorridos para um determinado alvo analisado e procura correlação entre os eventos.

Segundo Tang e Maclennan (2005, p.32, tradução nossa) “a principal diferença entre o modelo de seqüência e descoberta de associações é que no modelo de seqüência analisa o estado enquanto na associação analisa-se cada item”.

Pode-se imaginar um exemplo de um cliente que compra em uma loja um computador, após alguns dias volta e compra caixas de som e depois uma webcam o modelo de análise de seqüências irá analisar justamente eventos em série.

### Clusterização

De acordo com Goldschmidt e Passos (2005) a técnica de clusterização também conhecida como agrupamento é usada para particionar os registros de uma base de dados em subconjuntos. A clusterização se diferencia da classificação que na clusterização os rótulos são definidos automaticamente.

A análise de cluster envolve um conjunto de padrões previamente estabelecidos. O algoritmo irá analisar os conjuntos de dados, classifica-los e preestabelecer relações, é um método similar ao de classificação com a diferença de que não há tanta interferência humana.

## 5. ALGORITMOS DE DATA MINING

### K-Means

O objetivo do algoritmo K-Means é oferecer uma classificação de informações de acordo com os próprios dados. Pichiliani (2006) destaca que “essa classificação é baseada em análise e comparações entre os valores numéricos dos dados, o algoritmo irá fornecer uma classificação automática dos dados sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhuma pré-classificação existente, por essa característica K-Means é considerado como algoritmo de mineração não supervisionado”.

O Algoritmo K-Means pode ser considerado um método de clusterização, conforme explicado por Goldschmidt e Passos (2005) toma-se randomicamente, K pontos de dados como sendo os elementos do centro do cluster, em seguida cada ponto ou registro da base é atribuído ao cluster cuja distância deste cluster com relação ao centro do cluster é a menor dentre todas as distancias calculadas. Um novo centro é computado para cada cluster para a iteração seguinte, o processo termina quando o centro do cluster para de se modificar ou após um número limitado de iterações que tenha sido especificado pelo usuário.

## Apriori

O algoritmo Apriori é considerado um dos mais populares em relação a mineração com a técnica de regras de associação, esse algoritmo faz recursivas buscas no banco de dados a procura dos dados freqüentes. (MIRANDA, REIS, ARBEX, 2007).

Este algoritmo possui uma série de funções que são responsáveis por uma série de responsabilidades, funções de antimonotonia, função apriori-gen, função hash entre outras, o algoritmo apriori utiliza de consultas SQL[5] para buscar dados do banco de dados.

## Redes Neurais

De acordo com Goldschmidt e Passos (2005) algoritmos de redes neurais podem implementar as tarefas de classificação, regressão, previsão de series temporais e clusterização.

O algoritmo de redes neurais analisa os dados e tenta aprender sobre eles apresentando uma saída, o *back-propagation* algoritmo de redes neurais tenta aprender sobre o problema e minimizar a função de erro entre a saída gerada pela rede neural e a saída desejada.

Segundo Tang e MacLennan (2005) os algoritmos de redes neurais são mais sofisticados que os de árvore de decisão, as redes neurais possuem um conjunto de nós (neurônios) que formam a rede, existem três tipos de nós, entrada, oculto e saída. Cada aresta liga dois nós a um peso. A direção da aresta representa o fluxo de dados durante o processo de previsão.

Cada nó é uma unidade de processamento, os nós de entrada formam a primeira camada da rede, “cada nó de entrada é mapeado como uma entrada de atributo como sexo, idade ou rendimentos [...] os nós de saída representam os atributos previsíveis, podendo ser uma ou múltiplas saídas” (TANG; MACLENNAN, 2005, p.249, tradução nossa).

A (Figura 03) representa um exemplo da disposição do algoritmo de redes neurais.

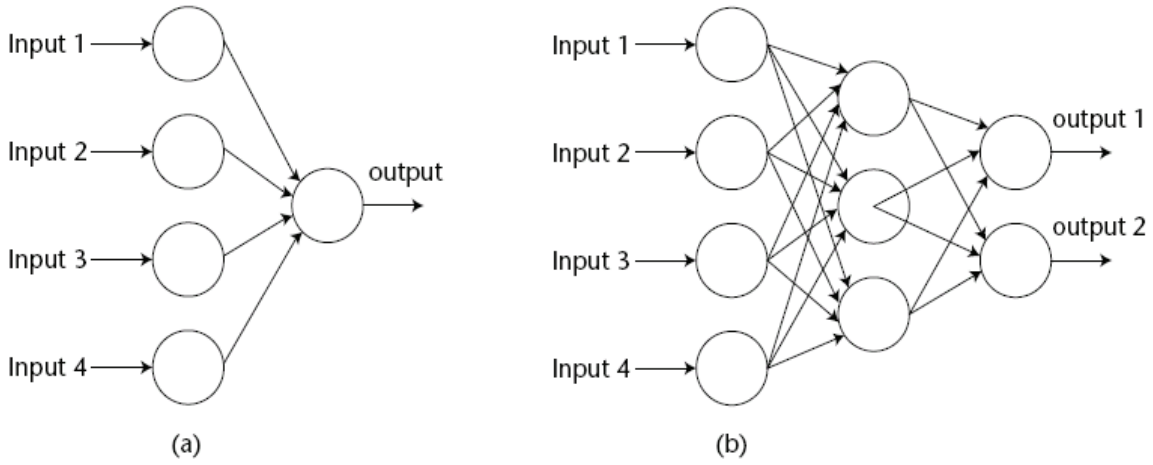


Figura 03: Exemplos de redes neurais
Fonte: TANG; MACLENNAN, 2005, p.250

[5] SQL: *Struct Query Language* é uma linguagem de pesquisa para banco de dados relacional

A (Figura 03a) mostra um exemplo de uma rede neural com 4 nós de entrada e um de saída, não há camada escondida doas as entradas se conectam a camada de saída. Na (Figura 03b) temos um exemplo de rede neural de três camadas, entrada, escondida e saída. Cada neurônio da camada entrada é conectado a camada escondida. Para Tang e MacLennan (2005, p.250) “a camada escondida é importante pois permite a rede aprender com relacionamentos não linear”.

Ainda citando Tang e MacLennan (2005, p.250) o processo de formatação envolve encontrar o maior conjunto de pesos para as arestas na rede e gerar previsões na camada de saída, baseado na configuração da rede o algoritmo calcula o erro para as realizações, com base nesses erros ele utiliza o *back propagation* para ajustar os pesos.

Nas redes neurais não há uma codificação de programas a fim de introduzir conhecimento sobre um problema. Por um processo iterativo (processo de aprendizado) as redes neurais lêem os exemplos fornecidos sobre um problema e criam assim um modelo de resolução. Elas são bem adaptadas a dois tipos de tarefas: reconhecimento de formas e generalização (ALMEIDA, 1995, apud GONÇALVES 2001, p.33).

## Árvore de decisão

O algoritmo C4.5 é um dos mais tradicionais algoritmos de classificação ele procura abstrair árvores de decisão a partir de uma abordagem recursiva de particionamento de bases de dados.

A idéia do algoritmo de árvore de decisão é montar uma estrutura onde cada nó da árvore representa uma decisão. Para Pichiliani (2006) a árvore é estrutura com três tipos de nó, o raiz que identifica o topo da árvore, os nós comuns que identificam um determinado atributo e os nós folha que contém as informações de classificação do algoritmo.

No algoritmo de árvore de decisão cada nó representa uma decisão sobre um atributo que determina como os dados estão particionados pelos nos filhos. (GOLDSCHMIDT; PASSOS, 2005)

Na (Tabela 08) apresenta-se um paralelo entre as técnicas e os algoritmos que as implementam, de acordo com os autores citados acima.

| Técnica | Algoritmos |
|---|---|
| Classificação | Árvore de decisão ou Redes Neurais |
| Regressão | Redes Neurais |
| Previsão | Redes Neurais |
| Descoberta de Associações | Apriori |
| Analise de seqüências | Árvore de decisão ou Redes neurais |
| Clusterização | K-means e Redes Neurais |

Tabela 01: Associação entre técnicas e algoritmos de *data mining*
Fonte: Arquivo pessoal

# 6. METODOLOGIA

Após o estudo da literatura apresentado anteriormente, procedeu-se a trabalho de aplicação de técnicas de mineração de dados a uma base de dados de uma Universidade onde selecionados dados do processo seletivo de alguns cursos.

O cadastro dos candidatos é feito em sistema próprio, é utilizado como pesquisa, para relatórios quantitativos e para análises superficiais. Não se tem um pensamento de utilizar essas informações para mineração de dados, a mineração de dados poderia auxiliar na identificação de prováveis inadimplentes, na avaliação de professores e matérias, na concessão de bolsas e outras atividades.

Utilizamos dados de um cadastro de candidatos a um processo seletivo em uma unidade de uma universidade particular, os dados foram extraídos do banco de dados que é alimentado pelo sistema acadêmico e importados para um estrutura que viabilizasse um grande volume de processamento.

# 7. RESULTADOS

Visando exemplificar a importância da mineração de dados aplicamos as técnicas e algoritmos de mineração de dados em uma base de dados preparada com os dados do primeiro processo seletivo de uma Universidade, analisando dados de uma unidade, que contém cursos de administração, computação, ciências contábeis, turismo e negócios internacionais.

Consideram-se as seguintes variáveis: região, sexo, turno, curso e aprovação, ou seja, se passou ou não passou. Os dados foram retirados do banco de dados que é alimentado pelo sistema acadêmico e importados para uma base de dados preparada para a mineração de dados.

Antes de importar os dados para a base preparada para a mineração foi realizada uma primeira etapa de seleção, retirando campos com cadastro errado ou campos nulos, antes de realizar a importação seleciona-se e substitui o campo bairro pela região respectiva, já que o sistema atual não oferece este dado, uma sugestão a implementar ou adquirir uma ferramenta dos correios que associa o CEP a região do cadastro.

Após a etapa de tratamento os dados foram importados para uma base de dados utilizando o sistema de gerenciamento de banco de dados SQL Server 2005, realiza-se o pré–processamento para retirar campos com erros na importação.

Na (Tabela 02) apresenta-se parte da tabela importada após a seleção e pré-processamento.

| | SEQ | CURSO | TUR... | SEXO | BAIRRO | PASSOU |
|---|---|---|---|---|---|---|
| 483 | 484 | Administracao | N | M | centro-sul | não |
| 484 | 485 | Computacao | N | F | centro-sul | não |
| 485 | 486 | Administracao | M | M | centro-sul | sim |
| 486 | 487 | Administracao | N | M | leste | não |
| 487 | 488 | Negocios internacionais | N | F | leste | não |
| 488 | 489 | Administracao | N | M | leste | sim |
| 489 | 490 | Computacao | N | M | leste | sim |
| 490 | 491 | Contabeis | N | F | leste | não |
| 491 | 492 | Administracao | N | M | centro-sul | não |
| 492 | 493 | Administracao | M | F | centro-sul | sim |
| 493 | 494 | Computacao | M | M | centro-sul | sim |
| 494 | 495 | Administracao | N | M | centro-sul | sim |
| 495 | 496 | Administracao | M | F | centro-sul | sim |
| 496 | 497 | Computacao | N | F | centro-sul | sim |
| 497 | 498 | Negocios internacionais | N | F | centro-sul | não |

Tabela 02: Parte da base de dados do 1º processo seletivo de 2007
Fonte : Arquivo pessoal

Utilizando a ferramenta de *Business* Intelligence da Microsoft o *SQL Business Intelligence Development Studio*, cria-se um *data source view*, que tem a mesma função de uma *view,* cria uma visão da tabela (Figura 04).

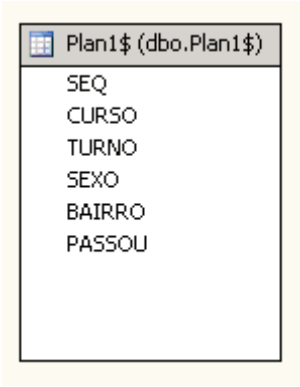


Figura 04: *Data Source view* obtido da tabela do processo seletivo
Fonte: Arquivo pessoal

De posse dos dados aplicam-se os algoritmos e técnicas de mineração, para o estudo de caso em questão destacam-se os algoritmos de redes neurais e árvore de decisão.

O algoritmo de árvores de decisão irá analisar e classificar os registros e nos propor algumas análises, já o algoritmo de redes neurais por abranger o maior número de tarefas nos apresentará previsões, classificações e probabilidades.

## Utilizando o algorítimo de Redes Neurais

Aplicando o algoritmo de redes neurais a base de dados pode-se retirar uma série de informações como:

- Em quais regiões tem-se maior impacto aos candidatos aos cursos de Administração e Computação?
- Qual o sexo que predomina para esses cursos?
- Qual o turno com maior número de candidatos?

De acordo com Tang e MacLennan (2005, p.262) autores do manual da Microsoft “a maior parte da tela mostra o impacto do atributo pelos pares de valores relacionados com um estado previsível”.

Pode-se dizer então que as marcas azuis destacam os valores mais importantes para os atributos relacionados.

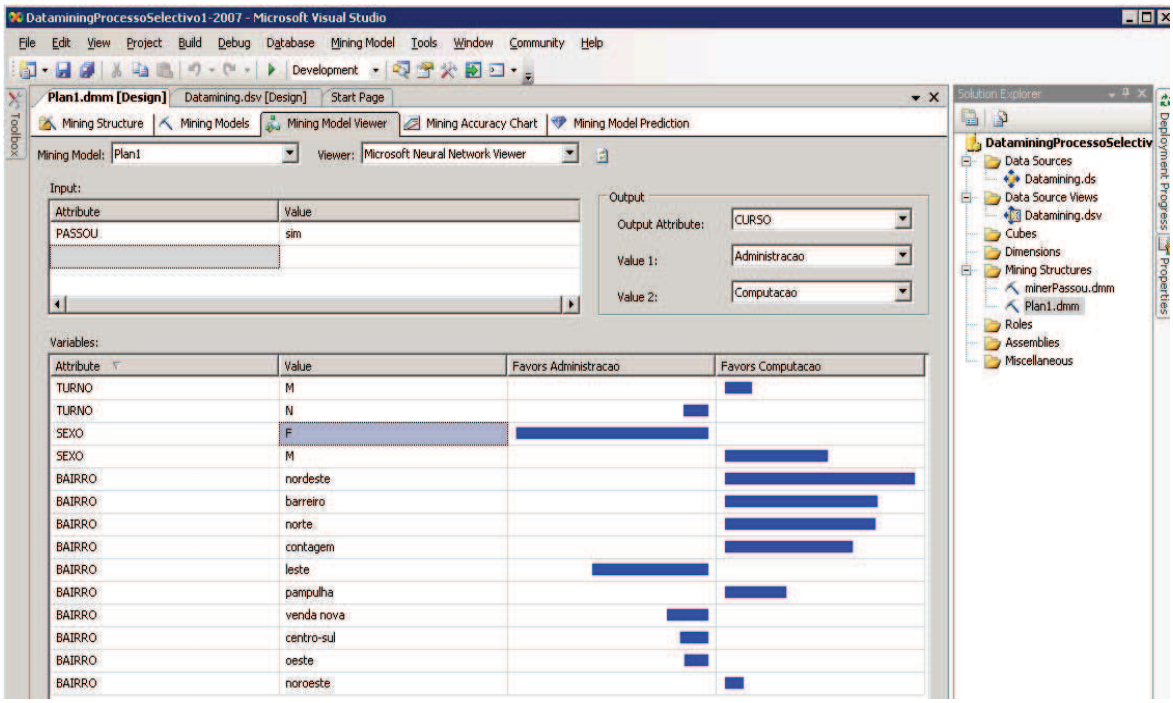


Figura 05: Análise 1 com algoritmo de redes neurais
Fonte: Arquivo pessoal

Na (Figura 05) percebe-se que para os alunos que passaram no vestibular para o curso de computação com relação ao atributo sexo o que mais pesa é o masculino. As principais regiões apontadas pelo algoritmo são as regiões nordeste e barreiro para o curso de computação, leste, venda nova e centro-sul para o curso de administração. O Algoritmo realiza uma série de análises e tenta relacionar as diversas variáveis da tabela e apontar as regiões propicias para alunos de computação e administração, essa hipótese merece uma análise detalhada e pesquisa para ser comprovada.

Filtrando ainda pelos alunos de administração e computação que não passaram no vestibular e fizeram inscrição para o turno da manhã, tem-se a seguinte situação, (Figura 06).

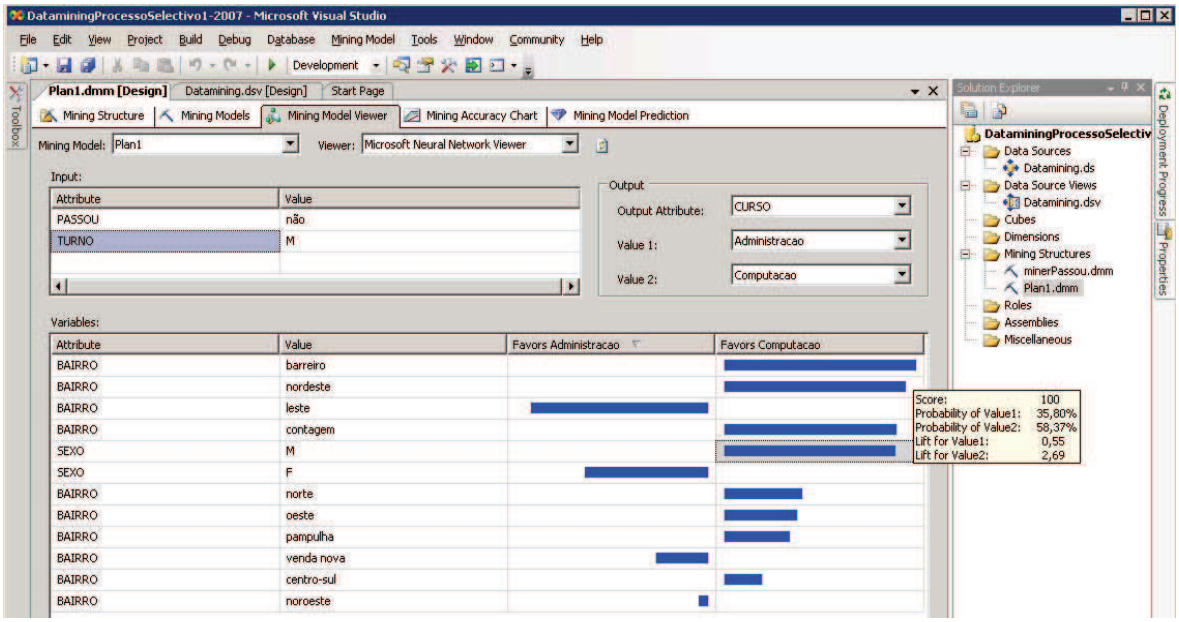

Figura 06: Análise 2 com algoritmo de redes neurais
Fonte: Arquivo pessoal

A região barreiro foi destacada pelo algoritmo de redes neurais para os alunos que optaram pelo curso de computação, já para o curso de administração a região leste, candidatos homens tendem a preferir computação e mulheres administração.

Estes resultados explicam o crescente investimento da Pontifícia Universidade Católica e da faculdade Novos Horizontes nessas regiões.

De acordo com Kotler (1998, apud Gonçalves, 2001) ao examinar repetidamente milhares de registros de dados, o software pode desenvolver um modelo estatístico poderoso descrevendo os relacionamentos e os padrões de dados importantes – nada que um pesquisador humano tenha tempo ou capacidade visual de fazer de maneira rigorosa e consistente.

## Utilizando o algoritmo Árvore de decisão

Aplicando a mesma tabela e condições do algoritmo de redes neurais para o algoritmo de árvore de decisão, segue-se o seguinte resultado (Figura 07).

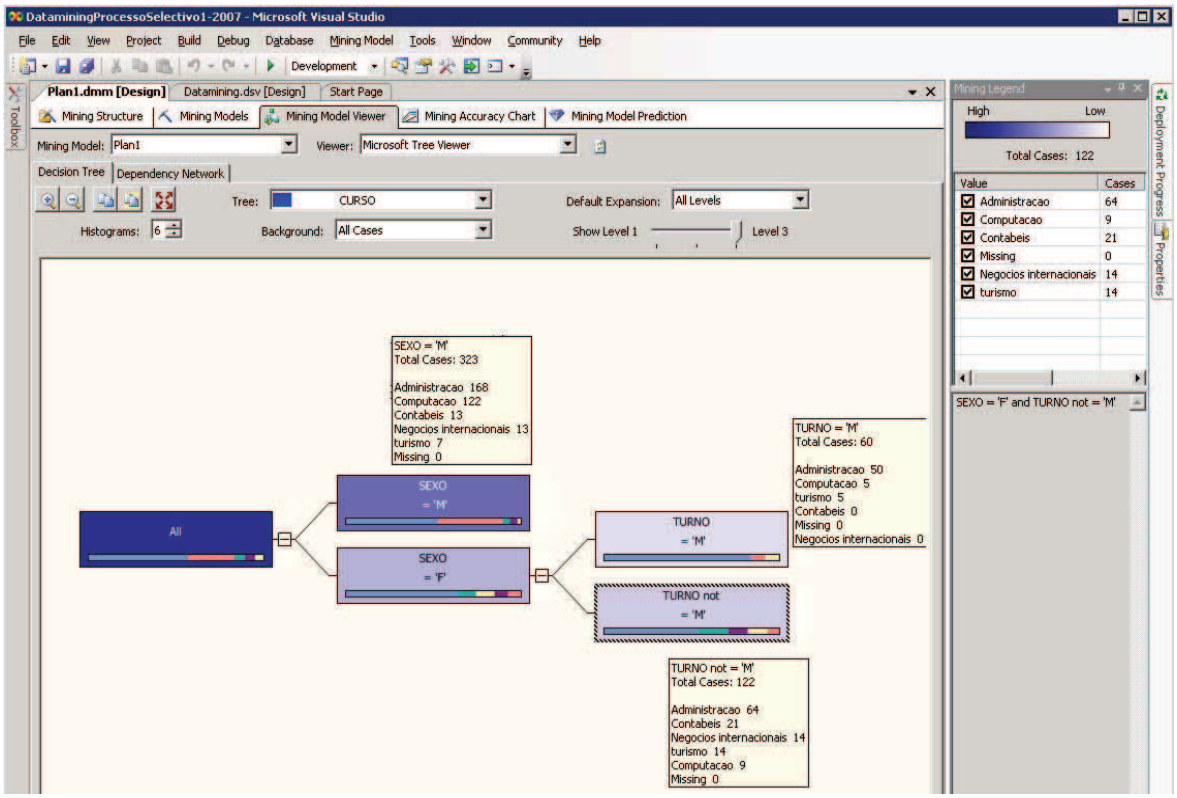


Figura 07: Análise 3 com algoritmo de árvore de decisão
Fonte: Arquivo pessoal

Percebe-se que o algoritmo de árvore de decisão não oferece a variedade de informações e hipóteses de previsão do algoritmo de redes neurais, mas oferece informações como: dos candidatos que prestaram vestibular para computação a maioria é do sexo masculino, 122 candidatos, do sexo feminino 14, sendo 9 para o turno da noite e 5 para manhã.

Na (Figura 08) destaca-se o curso da situação do curso de computação, neste curso passaram 64 candidatos sendo 33 para manhã e 31 para o turno da noite. Foram 72 candidatos reprovados sendo que 56 prestaram vestibular para o turno da noite e 16 para o turno da manhã, o que comprova que no turno da noite há maior concorrência.

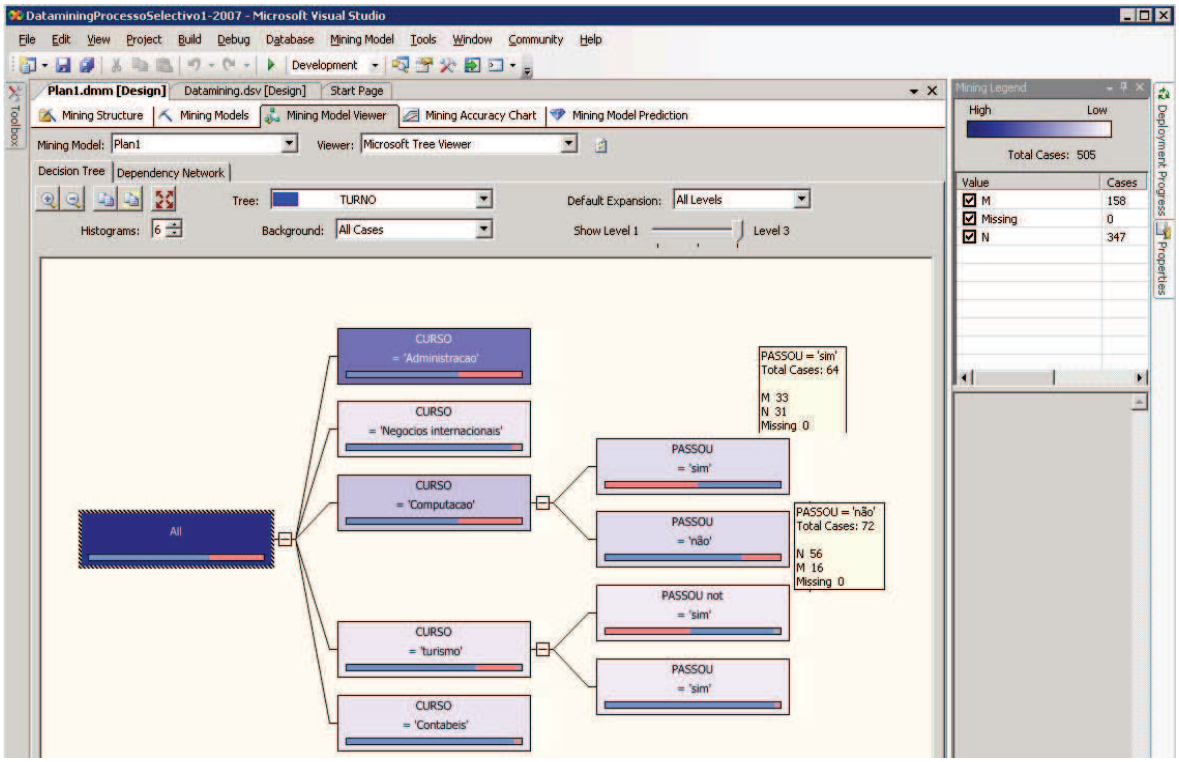

Figura 08: Análise 4 com algoritmo de árvore de decisão
Fonte: Arquivo pessoal

Comparando os dois algoritmos, árvore de decisão e redes neurais percebe-se que o algoritmo de árvore de decisão aponta uma hipótese especifica de acordo com o problema apresentado, já o algoritmo de redes neurais abrange uma maior variedade de modulação para tomada de decisão, onde ele estuda o problema com base os dados de entrada e tenta apresentar uma saída que trate uma variável de maior peso ou importância.

# 8. CONCLUSÕES

Procura-se ressaltar a importância do conteúdo que as empresas têm em suas bases de dados, destacam-se as diversas áreas e disciplinas que envolvem o processo de mineração de dados, evidenciando quais os conceitos e tecnologias necessárias para a iniciação de uma estrutura de mineração de dados como suporte a tomada de decisão.

Considera-se que a análise dados históricos é de extrema importância para as empresas compreenderem seu negócio, identificar futuros investimentos ou nichos de mercado, ou seja, oferecer informações que muitas vezes não são percebidas com análises superficiais.

O sucesso de implantação de uma estrutura de inteligência de negócios visando a mineração de dados depende de um projeto bem elaborado onde seja evidenciado os pontos que ele abrangerá e os problemas a se resolver. Na base a ser minerada deve-se realizar uma limpeza detalhada dos dados, reorganiza-los e após esta etapa aplicar os algoritmos e técnicas de *data mining.*

Um dos pontos principais da mineração de dados é parte de pré-processamento onde se realiza a seleção, organização e estruturação da base de dados, ao minerar dados sujos, ou incorretos as informações serão incorretas ou imprecisas.

Outro fator importante para o sucesso da mineração é que os usuários saibam qual informação desejam obter ou ao menos apontar um problema, é necessário que tenham algum conhecimento de informática para manusear a ferramenta e uma breve noção sobre a tecnologia de mineração de dados.

Algoritmos de mineração de dados pode-se ser aplicados a diversas situações visando associar, classificar, selecionar ou prever alguma situação não perceptível com uma análise superficial dos dados. É importante para o analista que venha a trabalhar com inteligência de negócios e mineração de dados tenha conhecimento dos algoritmos, que estão aplicando, saber sobre o funcionamento, particularidades e qual a melhor forma de organizar as variáveis a serem aplicadas, percebe-se que no estudo de caso que tivemos informações diferentes dependendo do algoritmo aplicado. Teve-se uma maior variedade de informações no algoritmo de redes neurais em relação ao algoritmo de árvore de decisão.

Confirma-se na prática a teoria levantada de que o processo de mineração de dados depende exclusivamente das três etapas, pré-processamento, mineração e pós-processamento destacando a importância de cada uma das etapas em especial a de pré-processamento que prepara os dados para a mineração.

No estudo de caso analisado percebe-se que o algoritmo de árvore de decisão classificou os candidatos do processo seletivo de acordo com a estruturação solicitada, por sexo e turno e por curso, onde se verificou que o curso de computação concentra se em sua maioria candidatos homens e a maior procura para o turno da noite, outros estudos podem aprofundar mais nesta relação entre sexo, turno e candidatos.

O algoritmo de redes neurais além de classificar como o de árvore de decisão, ele implementa fórmulas de previsão, no estudo de caso o algoritmo apontou as regiões barreiro e nordeste como regiões propicias para o crescimento, ou seja, que merecem um estudo especifico, confirma-se o potencial dessas regiões quando voltando para o mercado percebe-se que a Pontifícia Universidade Católica e outras faculdades como a Novos Horizontes estão investindo nessas regiões.

Conclui-se que o *data mining* pode e deve ser aplicado a qualquer situação problema, desde que possua-se dados para realizar uma análise detalhada, essas informações além de valiosas, são também decisivas para o processo de tomada de decisão.

Em um trabalho futuro pode-se apontar o uso da mineração de dados em conjunto com a inteligência competitiva, buscando não analisar somente os dados da própria empresa, mas também o conteúdo web de empresas concorrentes, associando essas informações as da própria empresa pode-se ter uma visão macro da situação a ser alcançada.

Quanto ao estudo de caso analisado como estudo futuro pode-se aplicar os conhecimentos de mineração de dados para analisar a região e seus alunos, ou até mesmo aplicar o data mining para identificar relações não percebidas com análises superficiais de relatórios gerenciais.

# 9. REFERÊNCIAS BIBLIOGRAFICAS

BARBIERI, Carlos. **Bi - business intelligence**: modelagem & tecnologia. Rio de Janeiro: Axcel Books, 2001. 424 p

_____. Notas de aula. BH, Universidade Fumec, aula de Banco de Dados II, abr. 2007.

BISPO, C.A. F **Uma analise da nova geração de sistemas de apoio a decisão.** São Carlos, 1998 143 p.

CARVALHO, I.C **Método de mineração de dados (data mining) como suporte a tomada de decisão.** 2002. 166 f. Dissertação (Mestrado em Ciência do Instituto Tecnológico da Aeronáutica) - São Jose dos Campos.

CARVALHO, R.B Administração de sistemas da informação. Belo Horizonte: Universidade Fumec, out, 2007. Notas de aula.

CHOO, Chun Wei. **A organização do conhecimento:** como as organizações usam a informação para criar conhecimento, construir conhecimento e tomar decisões. São Paulo,

CHU, S. Y.. **Banco de dados**: organização sistemas e administração, Editora Atlas. 1985, 398p.

DATE, C. J. **Introdução a sistemas de bancos de dados**. Rio de Janeiro, Campos, 1991. 674 p.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento empresarial:** como as organizações gerenciam o seu capital intelectual. Rio de Janeiro: Campus, 1998.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining:** um guia pratico: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005. 261 p.

GONÇALVES, L. P **Avaliação de ferramentas de mineração de dados como fonte de dados relevantes para a tomada de decisão:** aplicação na rede unidão de supermercados**.** 2001. 104 f. Dissertação (Mestrado em Administração) – Universidade Federal do Rio Grande do Sul. Porto Alegre.

GORDON, Steven R.; GORDON, Judith R. **Sistemas de informação:** uma abordagem gerencial. Rio de Janeiro: Livros Técnicos e Científicos, 2006. 377 p.

INMON, W.H. **What is a data warehouse?** v.1, n.1, 1997. Disponível em http://www.cait.wustl.edu/cait/papers/priscm/vol1_nol. Acesso em 22 jul. de 2007.

JAMIL, George Leal. **Gestão de informação e do conhecimento em empresas brasileiras**: estudo de múltiplos casos. Belo Horizonte: C/arte, 2006. 201 p.

JAMIL, George Leal. **Repensando a ti na empresa moderna**: atualizando a gestão com a tecnologia da informação. Rio de Janeiro: Axcel Books, 2001. 547 p.

MENDES, V.J. E FILHO E.E. **Sistemas integrados de gestão ERP em pequenas empresas:** um confronto entre o referencial teórico e a prática empresarial. n.3 São Paulo, dez 2002, p.277-296.

MIRANDA, D; REIS, D,B; ARBEX, E,C **Iniciação científica data mining.** Rio de Janeiro Faculdade Dom Bosco, 2003, Disponível em http://www.inf.aedb.br/datamining/index.html Acesso em 31 out. 2007.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro, 1997. 358 p.

PIACHILIANI, M. **Data mining na prática:** árvore de decisão. 2006. Disponível em http://www.imasters.com.br/artigo/5130/sql_server/data_mining_na_pratica_arvores_de_decisao/ Acesso em 31 de out. 2007

RESENDE, Solange. **Mineração de dados,** XXV Congresso da sociedade brasileira de computação, São Leopoldo, Rio Grande do Sul, UNISINOS, 2006. 397p.

SCHENATZ, B. N **Utilização de data mining em um sistema de informação gerencial para o diagnóstico da formação de professores da graduação.** 2005. 102 f. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina. Florianópolis.

SETZER, V **Dado, informação, conhecimento e competência**. Datagrama Zero, v.10, n.1, dezembro 2001. Disponível em http://www.dgz.org.br. Acesso em out. de 2007.