

Predição de Satisfação de Funcionários de uma Rede de Lojas do Comércio Varejista através da Aplicação de Técnicas em *Data Mining*

Autoria: Luiz Homero Bastos Cúnico, Marcia Aparecida Zampier

Resumo

O presente trabalho foi desenvolvido para prever a satisfação e insatisfação de funcionários de uma rede de lojas do comércio varejista, objetivando reduzir a rotatividade de pessoal, através de informações que permitam oferecer ao departamento de recursos humanos, subsídios no momento da contratação. Para tanto, utilizou-se de técnicas de predição em *Data Mining* (Mineração de Dados). Desenvolveu-se um estudo do processo e elaborou-se um resumo das diversas técnicas e de suas aplicações. Foram selecionadas três técnicas: Redes Neurais, Análise de Discriminante de Fisher e Regressão Logística. Aplicou-se um questionário aos colaboradores da empresa com perguntas relacionadas à satisfação e características pessoais de cada indivíduo. As respostas formaram o banco de dados que sofreu uma redução de dimensionalidade, através do Método das Componentes Principais e da Análise Fatorial e, na sequência, foi analisado nas três técnicas de *Data Mining*. Os resultados foram testados e avaliados fazendo-se um comparativo entre as técnicas e levando-se em consideração a margem de erro. No último estágio, um novo indivíduo foi avaliado. As conclusões identificaram que a técnica de Regressão Logística mostrou-se mais indicada para a tarefa proposta e os resultados podem servir de apoio na melhoria de futuras contratações.

1 Introdução

O ramo de venda a varejo é sem dúvida, um dos mercados mais competitivos que se tem notícia. A cada minuto, milhões de pessoas no mundo inteiro estão entrando em supermercados, lojas de conveniência, *drougstore*, etc, para realizar compras. No Brasil, o volume de capital envolvido é algo fabuloso, somente as três maiores redes de supermercados, Carrefour, Pão de Açúcar e Wal Mart faturaram no ano de 2007, cerca de R\$145.000.000,00 em média por dia (DIÁRIO OFICIAL DO NORDESTE).

Se o ramo de negócio envolve tanto dinheiro, significa que se pode almejar ótimos lucros. Porém, para entrar com competitividade neste mercado é necessário muito profissionalismo. Não basta atrair o cliente, com variedade de produtos, bons preços e lojas agradáveis, é necessário um algo a mais. Esse algo a mais é o atendimento personalizado que muitas vezes faz a diferença no momento de se optar por uma ou outra loja para adquirir produtos.

Um dos maiores problemas enfrentados atualmente dentro destas empresas é o que diz respeito à alta rotatividade de pessoal. Desta forma, o *Data Mining* (Mineração de Dados) foi utilizado na tentativa de descobrir padrões de comportamento, atitudes e características dentro do grupo de funcionários de algumas lojas de supermercados de uma rede, padrões esses, que evidenciem a satisfação e insatisfação dessas pessoas e subsidiem o departamento de recursos humanos com informações importantes no momento do recrutamento e seleção do pessoal contratado, minimizando a alta rotatividade.

Fica evidente que a rotatividade gera custos e quebra vínculos, porém o que fazer para evitá-la? Existe algo intrínseco, na personalidade das pessoas, que fazem com que elas se sintam satisfeitas ou insatisfeitas e fiquem mais tempo em seus postos de trabalho ou somente fatores externos influenciam nas rescisões de contrato de trabalho?

O objetivo principal do presente estudo é utilizar as técnicas de *Data Mining* para detectar padrões de comportamentos que provocam a satisfação e insatisfação dos funcionários de uma rede de supermercados e utilizar esses padrões para prever se um novo indivíduo pertence ao grupo dos satisfeitos ou insatisfeitos, e desta forma, diminuir a

rotatividade de funcionários.

Além da preocupação com a alta rotatividade, o trabalho também tem por finalidade a construção de um banco de dados com maior número de informações sobre os funcionários, objetivando futuras aplicações e descobertas.

Outro fator de suma importância diz respeito à análise dos dados processados pelos *softwares* matemáticos, que devem oferecer ao administrador de Recursos Humanos a leitura das informações de forma clara e aplicável à realidade da empresa, bem como auxiliar os diretores na tomada de decisão com base nas informações extraídas. Finalmente, se faz necessário acompanhar os resultados obtidos com o aprimoramento das técnicas, corrigindo erros e distorções.

2 Referencial Teórico

2.1 Recursos Humanos

Os departamentos de Recursos Humanos (RH) se preocupavam com a vida das pessoas e não com os negócios. Essa postura trouxe muitos obstáculos à evolução do Sistema de RH se comparado com outras áreas da empresa como a produção, o marketing, as finanças etc.

Segundo Carvalho e Nascimento (1999), a nova função do RH nas empresas atuais não deve estar voltada para o indivíduo tão somente, mas sim, à organização como um todo. Desta forma alguns pontos devem ser conferidos tais como: i) Ambiente Organizacional, englobando os desafios internos: estudos científicos das relações de trabalho e, externos: cultura ambiental, influência do mercado consumidor etc.; ii) Objetivos da organização: revisão de prioridades e metas da empresa; iii) Estrutura de cargos, responsabilidades e níveis de comando; iv) Motivação e liderança de equipes de trabalho; v) Relação de poder. A identificação desses itens é preponderante para que, o sistema de RH seja interpretado de forma moderna, dinâmica e arrojada.

De acordo com Chiavenato (1999), Boog e Boog (2002), Marras (2003), e Ribeiro (2005), as políticas de RH (agregar, aplicar, recompensar, desenvolver, manter e monitorar pessoas) estão subordinadas à filosofia empresarial e devem possuir flexibilidade, adaptando-se aos objetivos organizacionais.

Se por um lado a filosofia da empresa é algo duradouro e estável, por outro, as políticas de RH são mutáveis, dinâmicas e dependem de fatores externos como: i) das reações do mercado; ii) da influência do Estado; iii) da estabilidade política, econômica e social do país. Esses fatores, somados com a estratégia da empresa formam um cenário mediante o qual se fixarão metas de RH a curto, médio e longo prazos.

Na implantação dessas metas se devem visar os seguintes propósitos:

- Estabelecer programas e incentivos que objetivam a manutenção do funcionário na empresa por mais tempo, diminuindo consideravelmente os custos com a administração de funcionários;
- Proporcionar maior e melhor flexibilização em matéria de recrutar, selecionar, treinar e avaliar o desempenho dos funcionários da empresa;
- Adequar a administração de cargos e salários à dinâmica do mercado de trabalho.

Segundo Carvalho e Nascimento (1999), os objetivos do sistema aberto de Recursos Humanos podem ser classificados da seguinte maneira:

- Objetivos Societários = Proporcionar à empresa um sentimento de responsabilidade, face os desafios e necessidades da sociedade através do pagamento correto e dentro do prazo de impostos, taxas e programas assistenciais. Caso contrário, a empresa sofre restrições à ação no meio onde atua sob várias

formas: sanções legais (fisco), imagem arranhada ou distorcida, boicote aos produtos e serviços, etc;

- Objetivos Organizacionais = Transformar a empresa em um efetivo instrumento de integração organizacional, ou seja, o setor de RH deve ser visto como uma agência prestadora de serviços especializados para toda a empresa;
- Objetivos Funcionais = Manter em nível adequado seus procedimentos em função das necessidades efetivas de mão-de-obra treinada, consciente e responsável;
- Objetivos Individuais = Oferecer assistência aos funcionários na consecução de suas metas individuais, na medida que a administração participativa tende a se expandir na organização.

Segundo Chiavenato (1999), para haver produtividade no trabalho é necessário que os empregados sintam que o trabalho é adequado às suas capacidades e que eles são tratados equitativamente. Como as pessoas passam a maior parte de suas vidas no trabalho, é preciso que se identifiquem com o mesmo, pois a felicidade na organização e a satisfação no trabalho são fortes determinantes do sucesso organizacional.

O autor cita ainda que empregados satisfeitos não são necessariamente os mais produtivos, mas os insatisfeitos tendem a se desligar da empresa (aumento da rotatividade), se ausentar frequentemente (aumento de absenteísmo) e, produzir com pior qualidade.

De acordo com o departamento de recursos humanos da empresa pesquisada, os problemas gerados pela rotatividade de pessoal são muitos, mas dois destacam-se como mais importantes:

- A troca rotineira de funcionários gera custos de recrutamento e seleção, treinamento, rescisões contratuais e maior índice de ações trabalhistas;
- A rotatividade afeta significativamente os clientes das lojas, que se sentem bem, quando são recebidos e atendidos por funcionários já conhecidos. Portanto, perde-se com a rotatividade o lado social e humano.

2.2 Data Mining

Data Mining é um processo analítico utilizado para explorar uma grande quantidade de dados à procura de padrões consistentes ou relações sistemáticas entre as variáveis, o qual se divide em três fases: i) exploração inicial, ii) verificação/validação, iii) desenvolvimento (aplicação do modelo em novos dados para gerar predições) (STASOFT, 2008).

Carvalho (2001) define *Data Mining* como a utilização de técnicas automáticas de exploração de grandes quantidades de dados para descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano.

Segundo Andreatto (2002), *Data Mining* é um campo interdisciplinar que envolve a estatística, a Inteligência Artificial e o Aprendizado de Máquina.

De acordo com Louzada Neto e Diniz (2002), as aplicações de *Data Mining* têm sido observadas em várias áreas do conhecimento, entre elas estão as finanças, a saúde, criminologia, sociologia, ecologia, saneamento básico, climatologia, atuaria, manufatura, telecomunicações, governo, controle de qualidade, educação, comércio, marketing e medicina.

2.2.1 Tarefas e Técnicas em *Data Mining*

De acordo com Fayyad, Shapiro e Smity (1996), descobrir padrões e tendências escondidos em grandes massas de dados não é processo trivial. Em mineração de dados esse processo envolve o uso de diversas tarefas e técnicas. As tarefas são classes de problemas, que foram definidas através de estudos na área. As técnicas são grupos de soluções, ou seja,

algoritmos para os problemas propostos nas tarefas. Cada tarefa pode apresentar várias técnicas e algumas técnicas podem ser utilizadas para solucionar tarefas diferentes. As principais tarefas são: i) Associação, ii) Classificação; iii) Estimativa (Regressão) e, iv) Clusterização (análise de agrupamentos-*cluster analysis*).

As diversas técnicas de *Data Mining* possuem, por sua vez, vários algoritmos utilizados na solução dos problemas, muitos desses algoritmos estão disponíveis em softwares matemáticos e estatísticos através de pacotes que possuem tarefas específicas de acordo com o problema que se deseja solucionar.

O quadro 1 apresenta o resumo das tarefas em *Data Mining*, a função da tarefa, a técnica utilizada e a ferramenta (*software*) disponível para executá-la.

QUADRO 1 - CLASSIFICAÇÃO DAS TÉCNICAS EM DATA MINING

Tarefa	Função da Tarefa	Técnica	Algoritmos
Associação	Descrição	Regras de Associação	Apriori, DHP, ABS, Sampling, PARTITION, AQ, Lógica Fuzzi, WEKA
Classificação	Predição	Colônia de Formigas	Ant-Miner
		Árvores de Decisão	ID3, ID4, J48, CART, C4.5
		Algoritmo Genético	DeCan, RAGA
		Regressão Logística	Mínimos Quadrados, Máxima Verossimilhança
		Classificação Bayesiana	Teorema de Bayes (Probabilidade Condicional)
	Sumarização	Visualização	IBM, SAS, AVS, Visualization Edition
Estimativa	Predição	Redes Neurais	(MLP, Perceptron, Adelinee Madeline, Backpropagation, TANAGRA,
		Regressão Linear	Regressão estatística (M5, CART)
		Análise de Discriminante	Função Discriminante de Fischer
Clusterização	Descrição	Análise de Cluster	K-means, HAC

FONTE: Elaborado pelos autores (2005).

Neste estudo, serão apresentados conceitos e dados sobre a tarefa estimativa, cuja função é preditiva e, as técnicas relacionadas: Redes Neurais, Regressão Linear (Logística) e Análise de Discriminante.

A Tarefa Estimativa objetiva definir um valor numérico de alguma variável desconhecida a partir dos valores de variáveis conhecidas. A tarefa, como o próprio nome diz, se refere aos cálculos matemáticos. Exemplo de aplicação: estimar quantos carros passam em determinado pedágio, coletando informações como: cidades mais próximas, preço do pedágio, dia da semana, rodovia em que está localizado o pedágio, entre outros. Já a função preditiva, utiliza-se da resposta encontrada na Tarefa Estimativa, que geralmente é um número Real, compreendido entre zero (0) e um (1) e prediz se o indivíduo em estudo pertence a um determinado grupo ou a outro (VIANA, 2004).

2.2.1.1 Redes Neurais

De acordo com Wang (1995), Redes Neurais são uma classe de modelagem de prognóstico, que trabalha por ajuste repetido de parâmetro. Estruturalmente uma rede neural consiste em um número de elementos interconectados (chamados neurônios), organizados em camadas que aprendem pela modificação da conexão que une firmemente as camadas.

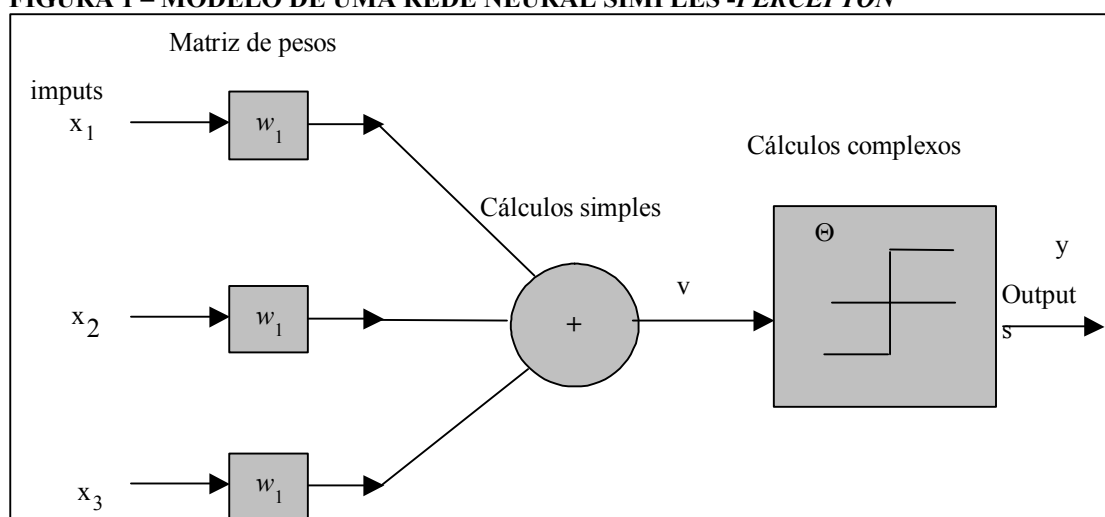
As Redes Neurais geralmente constroem superfícies equacionais complexas, através

de iterações repetidas, cada hora ajustando os parâmetros que definem a superfície. Depois de muitas repetições uma superfície pode ser internamente definida, pois se aproxima muito dos pontos dentro do grupo de dados (CARVALHO, 2002).

Cada vez mais a Teoria de Redes Neurais Artificiais consolida-se como um novo paradigma para abordagem da ampla classe dos assim chamados problemas complexos, em que extensas massas de dados devem ser modelados e analisados em um contexto multidisciplinar, envolvendo, simultaneamente, tanto aspectos estatísticos e computacionais como dinâmicos e de otimização. Os currículos de graduação em engenharia e ciências exatas estão sendo, neste momento, revistos para incluírem uma exposição introdutória de teoria de redes neurais artificiais, tema tratado já algum tempo em cursos de pós-graduação (KOVÁCS, 2002, p. 9-10).

Segundo Coutinho (2008), a função básica de cada neurônio é avaliar os valores de entrada (*inputs*), calcular o total para valores de entrada combinados, comparar o total com um valor limiar e determinar o valor de saída (*outputs*). Enquanto a operação de cada neurônio é bastante simples, procedimentos complexos podem ser criados pela conexão de um conjunto de neurônios, tipicamente, as entradas dos neurônios são ligadas a uma ou mais camadas intermediárias (denominadas camadas escondidas), que é então conectada com a camada de saída, conforme demonstrado na figura 1.

FIGURA 1 – MODELO DE UMA REDE NEURAL SIMPLES -PERCEPTON



FONTE: Kovács (2002, p. 30).

2.2.1.2 Regressão Logística

Quando se deseja estudar as relações entre um conjunto de variáveis independentes com uma variável dependente considera-se um problema multivariado. Na análise de tal problema, geralmente usa-se algum modelo matemático para lidar com as complexas inter-relações entre as diversas variáveis (MARQUES, 2004).

A regressão logística é a abordagem de modelagem matemática usada para descrever a relação entre diversas variáveis independentes e uma variável dependente dicotômica (0,1), diferentemente do modelo de regressão linear que possui resposta com variável contínua.

2.2.1.3 Análise de Discriminante

Segundo Louzada Neto e Diniz (2002), análise de discriminante é uma técnica estatística apropriada para discriminação e classificação. O princípio fundamental é derivar uma regra que possa ser utilizada para classificar de forma otimizada uma nova observação em uma classe previamente identificada.

Os objetivos imediatos desta técnica envolvem a descrição gráfica ou algébrica, das características diferenciais das observações de várias populações, além de classificar as observações em uma ou mais classes predeterminadas. A idéia é derivar uma regra que possa ser usada para classificar de forma otimizada uma nova observação a uma classe já rotulada.

A Análise de Discriminante é adequada nas situações onde se pretende separar duas ou mais classes de objetos, pessoas, clientes, empresa, produtos entre outros, ou alocar um novo objeto a uma das classes existentes, ou ambas as coisas.

O vetor $\underline{X}$ contendo variáveis 'x' possui componentes, tais como: idade, altura, peso, estado civil, etc. que são analisadas na tentativa de se buscar uma regra de classificação para um futuro indivíduo.

O algoritmo utilizado é a Função Discriminante de Fisher (FISCHER, 1938). Na aplicação da função duas populações são previamente definidas e denominadas de π_1 e π_2 . A função é construída sem assumir a existência de uma função, probabilidade associada a cada grupo.

3 Metodologia

Após análise de diversas técnicas de predição, as técnicas usadas foram: i) Técnica de Redução: Análise Fatorial e, ii) Técnicas de Predição: Redes Neurais Artificiais; Análise de Discriminante de Fischer e Regressão Logística, que foram escolhidas, por se adequarem melhor ao banco de dados disponível e oferecerem uma margem de erro aceitável, além de propiciarem rápido processamento computacional e leitura de dados acessíveis para usuários leigos. Além da escolha das técnicas, também foi realizada uma Análise Fatorial, técnica multivariada, utilizada para reduzir a dimensionalidade do problema, através do *software STATÍSTICA*.

Neste trabalho, a técnica de Redes Neurais, foi realizada utilizando-se o *software MATLAB*, que possui funções preparadas (*default*) para a efetivação da técnica. Pesquisa realizada nos manuais do *MATLAB* indicou que os comandos utilizados trabalham com o algoritmo *Backpropagation*, acrescido de outro algoritmo, denominado *Levenberg-Marquadt*, utilizado para correção do erro.

Um questionário foi aplicado com o maior número de funcionários possíveis e com várias perguntas (variáveis) que fizeram surgir o banco de dados inicial utilizado no trabalho.

Como o objetivo principal do trabalho foi prever se um determinado funcionário está ou não satisfeito com o seu emprego, as perguntas foram divididas em dois grupos: as relacionadas com a satisfação ou insatisfação dos funcionários e as relacionadas com as características pessoais de cada indivíduo.

Para facilitar o tempo de resposta e também a compilação dos dados, optou-se por perguntas que proporcionassem respostas objetivas e numéricas, com fácil visualização e que não deixassem o pesquisado em dúvida e nem tampouco desinteressado pelo assunto. Optou-se por elaborar 30 perguntas no total, dispostas em somente duas folhas de papel no formato A4, contendo 10 questões relacionadas com satisfação ou insatisfação e mais 20 relacionadas às características pessoais.

Depois de concluído, o questionário foi aplicado em 14 lojas de uma mesma rede de supermercados com um total de 826 funcionários. De posse das respostas dos 826 questionários, o próximo passo foi adaptar as respostas de forma que pudessem ser analisadas através das técnicas de *Data Mining*.

As respostas relacionadas à satisfação dos funcionários foram lançadas em uma planilha eletrônica de Excel para que os valores pudessem ser somados e nessa fase foi desenvolvida uma escala. De acordo com o resultado do somatório das respostas de cada funcionário foi definido se o funcionário estava satisfeito, parcialmente satisfeito ou

insatisfeito com o seu trabalho. Sendo assim, funcionários que alcançaram somatório até 16 pontos responderam entre ótimo e bom, portanto, estão satisfeitos. Funcionários cujo somatório ficou entre 17 e 25 responderam entre bom e regular, portanto, estão parcialmente satisfeitos. Finalmente, funcionários cujo somatório apresentou valor superior a 26 pontos tiveram concentração de respostas entre, ruim e péssimo, foram classificados como insatisfeitos, conforme figura 2.

FIGURA 2 – ESCALA CLASSIFICATÓRIA

$\text{somatório} \leq 16 \Rightarrow \text{satisfeitos}$
 $17 \leq \text{somatório} \leq 25 \Rightarrow \text{parcialmente satisfeitos}$
 $\text{somatório} \geq 26 \Rightarrow \text{insatisfeitos}$

FONTE: Elaborado pelos autores (2005).

Posteriormente, os funcionários cujo somatório apresentaram repostas de parcialmente satisfeitos e insatisfeitos foram unidos em um só grupo e receberam resposta 0 (zero) enquanto que os satisfeitos receberam resposta 1(um), dando característica binária a cada pesquisado. O quadro 2 mostra uma parte deste somatório, a escala de resultados e a classificação binária final.

QUADRO 2 - RESPOSTAS DAS PERGUNTAS DE SATISFAÇÃO E A CLASSIFICAÇÃO BINÁRIA

FUNCION.	18	19	20	21	22	24	25	28	29	30	SOMA	CLASSIFICAÇÃO	BINÁRIO
1	2	1	2	2	6	2	1	2	2	1	21	PARC. SATISFEITO	0
2	1	1	1	1	1	3	1	2	2	1	14	SATISFEITO	1
3	2	1	1	1	1	3	1	2	1	1	14	SATISFEITO	1
4	1	1	1	1	1	2	1	1	1	1	11	SATISFEITO	1
5	1	1	3	3	3	3	2	1	2	1	20	PARC. SATISFEITO	0
6	4	2	2	6	6	3	2	3	2	1	31	INSATISFEITO	0
7	2	1	3	2	2	2	2	2	1	1	18	PARC. SATISFEITO	0
8	2	2	1	2	2	3	1	2	2	1	18	PARC. SATISFEITO	0
9	2	2	2	3	6	3	2	3	3	1	27	INSATISFEITO	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
825	1	1	1	1	1	2	1	1	1	1	11	SATISFEITO	1
826	2	2	1	1	1	2	1	1	2	2	15	SATISFEITO	1

FONTE: Pesquisa dos autores (2005).

Resolvido o problema da classificação binária, o próximo passo concentrou-se com a parte do questionário que envolvia as perguntas relacionadas às características pessoais de cada funcionário. Essa segunda parte do questionário é composta de 20 perguntas. As respostas foram lançadas em planilha eletrônica do Excel e formaram uma matriz 826 x 20, ou seja, 826 funcionários (linhas) e 20 respostas para cada funcionário (colunas), conforme quadro 3.

QUADRO 3 – RESPOSTAS 20 PERGUNTAS (CARACTERÍSTICAS PESSOAIS) MATRIZ BRUTA

NR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	23	26	27
1	1	1	1	1	1	4	1	10	1	2	7	37	19	6	6	4	1	1	2	8
2	1	2	5	1	1	4	1	5	1	2	7	11	57	1	4	4	1	1	5	7
3	2	1	1	1	6	5	1	11	1	3	7	11	57	8	7	4	2	1	7	8
4	1	1	1	0	1	4	1	2	1	4	2	8	152	2	7	4	1	1	1	8
5	2	2	2	2	3	5	1	3	3	2	7	8	52	1	5	4	3	1	7	3
6	2	2	1	1	2	5	1	3	4	5	7	8	52	1	3	1	2	1	7	5
7	4	2	2	4	1	5	1	8	9	1	7	22	74	4	4	4	3	1	7	2
8	2	1	2	2	1	4	2	10	1	4	7	28	7	4	4	4	4	1	6	4
9	1	1	1	0	7	5	1	7	1	2	4	11	57	3	3	2	1	1	5	2
10	1	1	1	0	1	6	1	7	1	4	6	33	170	5	3	2	1	1	7	3
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
820	1	1	1	0	1	6	1	2	1	2	3	8	52	5	4	4	1	1	5	3
821	1	1	1	0	4	4	1	12	3	3	1	8	52	3	6	4	1	1	3	1
822	1	1	1	0	2	5	1	3	4	4	4	33	170	5	6	4	2	1	6	3
823	4	1	1	1	1	5	1	10	3	4	3	28	46	4	8	4	7	1	7	3
824	3	1	2	2	6	5	1	1	3	6	7	23	4	11	6	5	1	5	7	8
825	5	2	1	2	3	5	1	3	1	6	7	28	96	6	8	4	5	2	6	5
826	2	1	2	0	1	5	1	7	1	2	7	28	46	7	8	4	3	2	7	3

FONTE: Pesquisa dos autores (2005).

De posse desta matriz bruta o problema seguinte passou a ser a filtragem das respostas e adequação das mesmas em formato útil para aplicação de técnicas de *Data Mining*. Nessa etapa as respostas foram analisadas uma a uma e algumas modificações foram realizadas, outras variáveis foram acrescentadas e retiradas, na tentativa de criar uma segunda matriz que foi denominada de matriz preparada.

Após todas as modificações terem sido concluídas a nova matriz modificada, ficou no formato 826 x 21, ou seja, 826 indivíduos com 21 componentes (características variáveis), conforme quadro 4.

QUADRO 4 – RESPOSTAS 20 PERGUNTAS (CARACTERÍSTICAS PESSOAIS) MATRIZ MODIFICADA.

	idac	sexo	esta	filho	seq.	esco	rel.	mês	anir	cor	folg	set	per	dist	con	regi	sala	resc	loja	cid	idh
nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1	1	1	1	1	4	1	10	1	2	7	3	6	6	4	1	1	8	2	3	77
2	1	1	1	0	7	5	1	7	1	2	4	2	3	3	2	1	1	2	2	3	77
3	1	2	1	1	4	5	2	8	3	2	7	1	3	2	1	1	1	3	2	3	77
4	3	2	2	0	1	5	1	2	3	4	4	1	5	7	4	5	1	7	2	3	77
5	1	1	1	0	3	5	1	2	1	2	7	2	5	3	2	1	1	8	2	3	77
6	1	1	1	2	6	4	1	7	1	2	1	2	4	5	4	1	1	8	2	3	77
7	4	2	2	0	2	4	1	11	3	6	7	1	9	4	1	2	2	2	2	3	77
8	2	1	2	1	4	5	1	10	1	3	7	1	5	6	4	2	1	8	2	3	77
9	1	1	2	4	1	4	1	8	1	5	7	1	4	4	4	1	1	3	2	3	77
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
821	4	1	1	0	5	7	1	10	8	2	1	3	6	4	5	3	7	8	14	1	86
822	1	1	1	0	4	4	1	12	3	3	1	1	3	6	4	1	1	1	14	1	86
823	4	1	1	2	1	5	1	10	3	4	3	3	4	8	4	7	1	3	14	1	86
824	3	1	2	2	6	5	1	1	3	6	7	1	11	6	5	1	5	8	14	1	86
825	5	2	1	0	3	5	1	3	1	6	7	3	6	8	4	5	2	5	14	1	86
826	2	1	2	0	1	5	1	7	1	2	7	3	7	8	4	3	2	3	14	1	86

FONTE: Pesquisa dos autores (2005).

A matriz modificada, apresentada no quadro 4 representa os dados lapidados e pode ser tratada como um *Data Warehouse*. A partir dela vários estudos podem ser realizados e técnicas de *Data Mining* podem ser aplicadas.

O próximo passo partindo da matriz preparada foi a redução de dimensionalidade, verificando as variáveis que possuem menor influência na explicação da resposta de predição.

Para implementação das técnicas de redução de dimensionalidade, utilizou-se a Análise Fatorial através do software *STATÍSTICA*, que possibilitou vários cálculos e conclusões.

Um dos cálculos realizados com o auxílio do *STATÍSTICA* foi encontrar os autovetores através da Análise das Componentes Principais, conforme indica o quadro 5, que demonstra a presença de 14 autovalores maiores que 0,8. A última coluna do quadro reforça a idéia que 14 variáveis são as principais, dentre as 21 iniciais, pois explicam aproximadamente 84% da variância total. As variáveis de menor importância foram identificadas analisando os autovetores, são elas: Idh, idade, condução, setor, tipo de loja, estado civil e nº de registros em carteira.

QUADRO 5 – AUTOVALORES E VARIÂNCIA ACUMULADA

	Autovalores	% total	Cumul.	Cumul.
	Eigenval	Variance	Eigenval	%
1	2,649200	12,615237	2,649200	12,615237
2	2,345010	11,166714	4,994210	23,781951
3	1,608360	7,658856	6,602569	31,440807
4	1,289053	6,138346	7,891622	37,579153
5	1,167076	5,557505	9,058698	43,136658
6	1,113182	5,300866	10,171880	48,437524
7	1,041843	4,961156	11,213723	53,398680
8	1,026792	4,889484	12,240514	58,288164
9	0,996464	4,745068	13,236979	63,033232
10	0,938044	4,466878	14,175023	67,500110
11	0,909172	4,329393	15,084196	71,829503
12	0,878550	4,183573	15,962746	76,013076
13	0,846106	4,029074	16,808852	80,042150
14	0,839537	3,997794	17,648388	84,039944

FONTE: Pesquisa dos autores (2005).

Outro dado de suma importância que o pacote de análise fatorial do *STATÍSTICA* oferece é o cálculo dos pesos fatoriais que permite distinguir qual fator dos 14 está carregado com quais variáveis. Para cálculo dos pesos fatoriais, optou-se por uma rotação dos fatores, também conhecida como rotação varimax, a rotação permite que os fatores sejam translacionados próximos de variáveis que os carregam com mais intensidade.

O quadro 6 mostra os pesos fatoriais, já rotacionados, cuja análise mostra claramente o carregamento do fator 1 pelas variáveis 1, 3 e 16 enquanto que o segundo fator está carregado pelas variáveis 19, 20 e 21.

QUADRO 6 – COEFICIENTES DOS ESCORES FATORIAIS ROTACIONADOS

	vr 1	vr 2	vr 3	vr 4	vr 5	vr 6	vr 7	vr 8	vr 9	vr 10	vr 11	vr 12	vr 13	vr 14	vr 15	vr 16	vr 17	vr 18	vr 19	vr 20	vr 21
ft 1	0,38	0,16	0,48	-0,01	-0,07	-0,02	0,03	0,02	-0,01	-0,06	-0,02	-0,01	-0,08	-0,02	-0,04	0,40	0,11	-0,05	0,01	0,00	0,01
ft 2	0,01	0,06	0,04	-0,02	-0,02	0,00	0,02	0,02	0,04	-0,01	-0,01	0,03	-0,02	0,12	0,02	-0,07	-0,06	-0,03	-0,34	0,40	-0,46
ft 3	0,01	-0,49	-0,16	-0,07	0,02	0,01	-0,01	-0,02	0,02	0,02	-0,04	0,51	-0,06	-0,08	-0,01	0,05	0,34	0,02	-0,09	-0,10	0,03
ft 4	-0,03	0,07	-0,02	-0,02	-0,04	-0,03	-0,02	-0,01	0,03	0,00	0,00	-0,02	0,00	0,63	0,54	-0,03	-0,02	-0,04	-0,14	-0,03	-0,06
ft 5	-0,04	0,00	0,02	0,03	-0,01	-0,02	-0,04	0,01	0,98	0,00	0,00	0,03	0,01	0,07	-0,01	0,01	-0,01	0,03	0,15	0,16	-0,06
ft 6	0,06	0,13	-0,12	0,00	0,98	-0,02	-0,04	0,01	-0,02	-0,05	0,00	0,07	-0,06	-0,05	-0,02	-0,08	0,05	-0,02	0,02	0,00	0,03
ft 7	-0,09	-0,01	0,11	-0,01	0,06	0,07	-0,02	0,01	0,00	0,03	0,02	0,13	-0,91	0,12	-0,12	0,14	-0,19	0,03	-0,14	-0,05	-0,01
ft 8	0,00	-0,11	-0,01	0,02	0,02	0,09	0,00	0,00	-0,02	0,01	-0,02	-0,02	0,04	0,10	-0,02	0,13	-0,12	-0,95	0,08	0,14	-0,01
ft 9	-0,02	-0,06	0,03	-0,02	0,01	0,03	0,01	1,00	0,01	0,04	0,00	-0,04	0,00	-0,02	0,00	0,01	-0,03	0,00	0,02	0,03	-0,02
ft 10	-0,04	-0,05	0,02	-0,01	0,00	0,06	-0,02	0,00	0,00	0,03	1,01	-0,06	-0,01	0,02	-0,02	-0,02	-0,05	0,02	0,02	0,00	0,01
ft 11	0,02	0,13	-0,10	-0,02	-0,05	-0,07	0,00	0,03	0,00	0,96	0,02	0,06	-0,04	-0,10	0,08	-0,04	0,06	-0,01	-0,06	-0,06	0,01
ft 12	0,02	0,01	0,05	-0,98	0,00	0,00	0,01	0,02	-0,02	0,01	0,02	0,10	0,00	0,12	-0,09	-0,03	-0,01	0,02	0,03	0,08	0,00
ft 13	0,03	-0,01	-0,13	0,00	0,04	0,00	-1,00	-0,01	0,04	0,00	0,02	0,02	-0,02	-0,02	0,06	0,07	-0,02	0,00	0,04	0,01	0,02
ft 14	0,00	-0,41	-0,15	0,00	0,00	-0,76	-0,01	-0,02	0,04	0,06	-0,05	-0,07	0,07	0,15	-0,09	0,09	-0,40	0,06	-0,01	-0,02	-0,03

FONTE: Pesquisa dos autores (2005).

4 Análise dos Resultados

Após a aplicação dos questionários aos funcionários e da tabulação dos resultados via Excel, foi efetuada a aplicação das três técnicas de *Data Mining* e o cálculo dos pesos e bias nas Redes Neurais, coeficientes de Fisher e os parâmetros estimados para o modelo logístico, para então, fazer a predição de novos indivíduos, identificando em que grupo eles pertenceriam.

Para realizar a predição, primeiramente deve-se aplicar o questionário para os novos indivíduos, somente com as perguntas relacionadas às características individuais, ou seja,

vinte e uma perguntas.

As respostas dessas perguntas, para cada novo indivíduo, formarão um vetor, primeiramente 1×21 , que deverá ser transposto e normalizado utilizando-se as médias e os desvios padrões de cada variável original, previamente calculada. De posse das médias e dos desvios, transforma-se o novo vetor com 21 variáveis ' x_i ' em um novo vetor de mesma dimensão, porém, com variáveis padronizadas ' z_i '. A padronização é feita através da equação 1.

$$z_i = \frac{x_i - \bar{x}_i}{s_i} \quad (1)$$

Onde:

x_i é a i-ésima resposta da variável original.

$\bar{x}_i$ é a média da i-ésima variável (calculada com as 826 variáveis originais).

s_i é o desvio padrão da i-ésima variável (calculado com as 826 variáveis originais).

O vetor com as variáveis normalizadas Z (21×1), equação (3) sofrerá agora a multiplicação pela matriz dos coeficientes dos escores fatoriais $(L'L)^{-1}L'$, gerada na análise fatorial (conforme quadro 6), através da equação (2), produzindo o vetor f_i no formato (14×1), ou seja, quatorze fatores para cada indivíduo conforme segue.

$$f_i = (L'L)^{-1}L'.Z \quad (2)$$

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ . \\ z_{21} \end{bmatrix} \quad (3)$$

Onde:

L é a matriz dos pesos ou carregamentos após a rotação varimax, gerado na análise fatorial.

L' é a matriz transposta de L

$(L'L)^{-1}L'$ é a matriz dos coeficientes dos escores fatoriais

Z é a matriz das variáveis padronizadas.

f é o novo vetor reduzido.

O quadro 7 mostra cinco novos candidatos ao emprego, com as 21 respostas do questionário, as médias e os desvios padrões das variáveis originais, as variáveis normalizadas e a redução para 14 fatores.

QUADRO 7 - CINCO NOVOS INDIVÍDUOS COM 14 FATORES

	Novos indivíduos					Média	d.p	Novos Ind. Normalizados					Redução para 14 fatores				
variável	x1	x2	x3	x4	x5	xm	s	z1	z2	z3	z4	z5	f1	f2	f3	f4	f5
idade	3	2	2	3	5	2,31	1,45	0,48	-0,21	-0,21	0,48	1,86	0,97	-0,23	-0,75	0,42	1,67
sexo	2	2	1	1	2	1,46	0,50	1,09	1,09	-0,92	-0,92	1,09	0,95	0,48	9,05	-0,40	-1,79
estado	2	1	1	2	2	1,65	0,77	0,45	-0,84	-0,84	0,45	0,45	0,61	-0,85	0,35	1,27	-1,13
filhos	2	1	0	2	3	0,83	1,10	1,06	0,15	-0,76	1,06	1,98	0,43	-1,27	0,33	0,63	-0,06
seq.	2	1	2	2	3	2,98	1,98	-0,50	-1,00	-0,50	-0,50	0,01	-0,54	-0,56	1,76	-0,45	-0,63
escol.	3	3	4	4	2	4,33	1,13	-1,18	-1,18	-0,29	-0,29	-2,06	-0,26	-0,65	-0,78	-0,52	0,14
rel.	1	1	1	2	1	1,51	1,56	-0,33	-0,33	-0,33	0,31	-0,33	0,64	-0,02	-0,10	0,10	-0,84
mês	6	1	9	4	11	6,49	3,42	-0,14	-1,60	0,73	-0,73	1,32	-1,52	0,07	0,62	0,79	-1,33
anima	1	1	3	1	1	2,22	2,15	-0,57	-0,57	0,36	-0,57	-0,57	-0,35	-1,72	1,14	-0,79	1,10
cor	3	2	2	4	2	3,69	1,77	-0,39	-0,96	-0,96	0,17	-0,96	-1,73	-0,79	-1,29	0,61	-1,25
folga	1	3	2	7	2	4,92	2,81	-1,39	-0,68	-1,04	0,74	-1,04	0,01	-0,71	-1,19	0,12	-0,72
set	3	1	2	3	1	1,61	0,80	1,74	-0,76	0,49	1,74	-0,76	-0,95	-0,26	0,90	-0,85	-2,10
perm	3	4	5	5	8	5,25	4,74	-0,47	-0,26	-0,05	-0,05	0,58	0,27	0,25	0,03	-0,29	0,45
dist.	5	3	4	7	5	5,30	2,16	-0,14	-1,06	-0,60	0,79	-0,14	0,15	0,70	0,99	0,18	1,34
cond.	4	1	2	4	4	3,22	1,33	0,59	-1,67	-0,92	0,59	0,59					
registro	4	2	2	3	5	2,59	1,62	0,87	-0,37	-0,37	0,25	1,49					
salário	2	1	2	2	1	1,40	1,06	0,57	-0,38	0,57	0,57	-0,38					
escisã	8	4	4	3	8	4,85	2,39	1,32	-0,36	-0,36	-0,78	1,32					
loja	3	7	9	11	13	8,33	3,76	-1,42	-0,35	0,18	0,71	1,24					
cid	3	3	4	2	1	2,65	1,18	0,30	0,30	1,14	-0,55	-1,40					
idh	77	77	7	80	86	80,03	3,81	-0,80	-0,80	-19,18	-0,01	1,57					

FONTE: Elaborado pelos autores (2005).

Esses novos vetores receberão os pesos e bias e a função *logit*, os coeficientes de Fisher e os parâmetros beta da regressão e retornarão a uma resposta numérica compreendida entre zero e um para as três técnicas, possibilitando a classificação do novo indivíduo como satisfeito ou insatisfeito.

Para realizar as aritméticas com o vetor 14 x 1, foi desenvolvida planilha do Excel e parte desta planilha com as respostas segue no quadro 8.

QUADRO 8 - RESULTADOS DAS PREDIÇÕES PARA CINCO NOVOS INDIVÍDUOS

novo indivíduo	redes	predição	fisher	predição	regressão	predição
y1	0,93054	satisfeito	1,56858	insatisfeito	0,9088056	satisfeito
y2	0,33170	insatisfeito	0,14221	insatisfeito	0,7067018	satisfeito
y3	0,66195	satisfeito	2,73057	insatisfeito	0,9680499	satisfeito
y4	0,57265	satisfeito	-0,1851	satisfeito	0,6477692	satisfeito
y5	0,37405	insatisfeito	0,40076	insatisfeito	0,7683513	satisfeito

FONTE: Elaborado pelos autores (2005).

A utilização das três técnicas em conjunto, aplicadas a cada novo candidato ao emprego, indica ao profissional de R.H. qual é o candidato ideal a ser contratado. Essa escolha não substitui o recrutamento e a seleção previamente realizados, mas sem dúvida, proporciona uma escolha com maior racionalidade, alicerçada em pesquisa contendo dados de funcionários satisfeitos e insatisfeitos.

A escolha racionalizada de novos contratados e a constante pesquisa e atualização do banco de dados, poderá proporcionar ao longo do tempo, refinamento das técnicas, minimizando as margens de erro e, sobretudo, permitindo a contratação de colaboradores com tendências à satisfação e conseqüentemente, reduzindo os níveis de rotatividade, objetivo

principal deste trabalho.

Assim sendo, após a execução do trabalho e analisando-se o quadro 8, pode-se concluir, com determinado grau de certeza que o indivíduo que deve ser contratado é o 'y4', pois nas três técnicas o resultado predito para esse indivíduo foi a satisfação.

5 Considerações Finais

A tarefa de prever satisfação e insatisfação de funcionários, não é nada simplista. A complexidade da tarefa vai desde o fato de que se está trabalhando com o ser humano e com suas subjetividades de opiniões e características individuais, até o conceito propriamente dito do que seja estar satisfeito e não estar satisfeito.

Diante dessa situação complexa o trabalho mostrou-se um desafio muito grande, mesmo por que, até o momento não se tinha conhecimento de técnicas de *Data Mining*, estarem sendo empregadas para prever satisfação e insatisfação de funcionários.

Os primeiros desafios surgiram na escolha das técnicas a serem aplicadas. Para desenvolver trabalhos utilizando-se técnicas de *Data Mining*, se faz necessário um conhecimento prévio e profundo, de cada técnica e do processo de Mineração de Dados propriamente dito. Durante este trabalho tomou-se conhecimento das mais variadas técnicas e seus métodos algoritmos e aplicações.

Outro fator decisivo foi a escolha do banco de dados. Após análise do banco de dados disponível na empresa e que naquele momento era pobre em informações, optou-se por construir um novo banco de dados entrevistando funcionários, através do questionário e esse banco de dados foi trabalhado, filtrado e lapidado, passo a passo até constituir-se em um novo *Data Warehouse*. Essa tarefa foi a mais árdua, pois foram despendidos meses de trabalho.

A escolha das três técnicas: Redes Neurais, Análise de Discriminante de Fischer e Regressão Logística, dentre tantas estudadas, foram pertinentes, visto que o problema caracterizava-se em uma classificação binária e qualquer uma dessas técnicas pode fazer a tarefa.

As técnicas de Análise Fatorial, através da Análise de Componentes Principais, foram essenciais para enriquecer o trabalho e fornecer conhecimentos na área de análise multivariada, que vêm sendo largamente utilizadas em trabalhos acadêmicos, com o advento do computador e de novos *softwares*.

O uso de algoritmos matemáticos possibilitou o conhecimento mais profundo do *software MATLAB e STATÍSTICA*, sendo que nas técnicas Redes Neurais, Fischer e Análise Fatorial, trabalharam-se com programação no *MATLAB*.

Os resultados conseguidos nas três técnicas foram razoáveis se considerado a complexidade da classificação. O destaque foi para a técnica Regressão Logística que apresentou o menor percentual de erro na fase de teste, seguido da Análise de Discriminante de Fisher e, finalmente, as Redes Neurais.

O administrador de Recursos Humanos pode agora utilizar os pesos e bias das Redes Neurais, os coeficientes de Fischer e os parâmetros estimados (betas) da Regressão Logística para prever, mediante as três técnicas, se um novo indivíduo, postulante a uma vaga de emprego, estaria no grupo dos satisfeitos ou insatisfeitos, guardadas as devidas margens de erro.

Este trabalho abre caminho para uma série de análises em *Data Mining*. Partindo-se do *Data Warehouse* elaborado e melhorado, outras técnicas podem ser utilizadas, entre elas merecem destaque as Árvores de Decisão e as Regras de Classificação. Outras análises podem ser feitas com a separação do grupo inicial, por loja (filial), por cidade, por número de funcionários etc.

Na tentativa de reduzir a margem de erro, o questionário inicial poderia ser

aumentado para 60 perguntas distribuídas entre satisfação pessoal e características pessoais. Esse aumento de variáveis deixaria o *Data Warehouse* mais robusto e as técnicas com mais consistência.

Outro fator importantíssimo e que merece destaque num futuro trabalho, é o formato das respostas no questionário. A substituição de números aleatórios, como foi feito, por opções que pudessem carregar pesos nas respostas, fariam com que as técnicas e estatísticas de *Data Mining*, tivessem mais significância, diminuindo o erro, visto que se trabalha muito com médias, variâncias, autovalores e autovetores, ou seja, conceitos estatísticos quantitativos.

Faz-se necessário também, concluir sobre as divisões dos dados em parcelas de treinamento com 70% e teste com 30%. Apesar de toda aleatoriedade dispensada nessa fase, algum vício e distorção pode ter surgido. Talvez fosse mais perspicaz elaborar um programa no próprio *MATLAB*, que procedesse todo o treinamento com os demais indivíduos e o primeiro indivíduo apenas sendo testado. Após deixar-se-ia para teste o segundo indivíduo e treinavam-se os restantes e assim sucessivamente até que o melhor resultado fosse obtido, porém, tal procedimento pode requerer um certo tempo computacional.

Finalmente, a utilização das técnicas e a atualização do *Data Warehouse* periodicamente pelo departamento de R.H são imprescindíveis para refinação dos dados e das respostas, que com o tempo, tendem a ficar cada vez mais confiáveis.

Referências

- ANDREATTO, R. **Construindo um *Data Warehouse* e Analisando suas informações com *Data Mining* e OLAP**. Faculdade de Ciências Administrativas. Valinhos/SP: 2002.
- BOOG, G.; BOOG, M. (Coord.). **Manual de gestão de pessoas e equipes: Estratégias e tendências**. v. 1. São Paulo: Gente, 2002.
- CARVALHO, A. P. de L. F. de. **Redes Neurais Artificiais**. 2002. Disponível em: <<http://www.icmc.sc.usp.br>>. Acesso em: 11/2004.
- CARVALHO, A. V. de; NASCIMENTO, L. P. **Administração de recursos humanos**. v.1. São Paulo: Pioneira, 1999.
- CARVALHO, L. A. **Data Mining: A mineração de dados no marketing, medicina economia, engenharia e administração**. São Paulo: Érica, 2001.
- CHIAVENATO, I. **Gestão de pessoas: O novo papel dos recursos humanos nas organizações**. Rio de Janeiro: Campos, 1999.
- COUTINHO, F. V. **Data Mining**. 2008. Disponível em: <<http://www.dwbrasil.com.br/html/dmining.html>>. Acesso em: 04/2008.
- DIÁRIO OFICIAL DO NORDESTE. **Carrefour lidera ranking de supermercados no País**. 2008. Disponível em: <<http://diariodonordeste.globo.com/noticia.asp?codigo=211751&modulo=968>>. Acesso em: 04/08.
- FISHER, R. A. **The statistical utilization of multiple measurements**. Annals of Eugenics, 1938.
- FAYYAD, U.; SHAPIRO, G; SMYTH, P. **Advances in Knowledge Discovery & Data Mining**. Cambridge, MA: AAAI/MIT, p. 1-34, 1996.

KOVÁCS, Z. L. **Redes neurais artificiais: Fundamentos e aplicações**. 3. ed. São Paulo: Livraria da Física, 2002.

LOUZADA NETO, F.; DINIZ, C. A. R. **Técnicas estatísticas em *data mining***. Monografias Del IMCA. Ano 31, Lima-Peru, 2002.

MARQUES, J. M. Notas de aula da disciplina de Análise Multivariada Aplicada à Pesquisa. Curso de Mestrado Interinstitucional em Métodos Numéricos em Engenharia. Fev/2004.

MARRAS, J. P. **Administração de Recursos Humanos: Do operacional ao estratégico**. 7. ed. São Paulo: Futura, 2003.

RIBEIRO, A. de L. **Gestão de pessoas**. São Paulo: Saraiva, 2005.

STASOFT. **Data Mining Techniques**. Disponível em: <<http://www.statsoft.com/textbook/stathome.html?stdatmin.html&1>>. Acesso em: 04/2008.

VIANA, R. Mineração de dados: Introdução e aplicações. **Revista SQL Magazine**. 10 ed. Ano 1. p. 27 e 28. Rio de Janeiro: Neofício, 2004.

WANG, S. **The Unpredictability of Standard Back Propagation Neural Networks in Classification Applications**. Management Science, 1995, v. 41, n. 3, p. 555-559.