LUIZ HOMERO BASTOS CUNICO

TÉCNICAS EM *DATA MINING* APLICADAS NA PREDIÇÃO DE SATISFAÇÃO DE FUNCIONÁRIOS DE UMA REDE DE LOJAS DO COMÉRCIO VAREJISTA.

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, Curso de Pós-Graduação em Métodos Numéricos em Engenharia – Programação Matemática, Setores de Tecnologia e Ciências Exatas, Universidade Federal do Paraná.

Orientador: Prof. Dr. Celso Carnieri.
Co-orientador: Prof Dr. Anselmo Chaves Neto

CURITIBA
2005

TERMO DE APROVAÇÃO

Luiz Homero Bastos Cunico

Técnicas em *Data Mining* Aplicadas na Predição de Satisfação de Funcionários de uma Rede de Lojas do Comércio Varejista.

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Curso de Pós-Graduação em Métodos Numéricos em Engenharia – Área de Concentração em Programação Matemática, Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: **Prof. Celso Carnieri**
Programa de Pós-graduação em Métodos Numéricos em Engenharia - UFPR

Co-orientador: **Prof. Dr Anselmo Chaves Neto**
Departamento de Estatística – UFPR

Prof. Dr Jair Mendes Marques
Programa de Pós-graduação em Métodos Numéricos em Engenharia - UFPR

Prof. Dr Celso Kaestner
Programa de Pós-graduação em Informática Aplicada-PUC-PR

Curitiba, 23 de setembro de 2005.

Dedico este trabalho ao meu pai Antonio Carlos (in memorian).
"*Seu exemplo de vida será sempre o meu caminho*".

AGRADECIMENTOS

Aos professores orientadores Celso Carnieri e Anselmo Chaves Neto, pelos ensinamentos e orientação, pela amizade e companheirismo.

Ao professor Jair Mendes Marques, pelas tantas vezes que prontamente me atendeu e contribui com preciosos ensinamentos e orientações.

Aos professores Liliana Madalena Gramani Cumin, Maria Teresinha Arns Steiner, Neida Maria Patias Volpi e Volmir Eugênio Wilhem, pelos ensinamentos recebidos durante o curso.

A Universidade Federal do Paraná e a Universidade Estadual do Centro Oeste pela iniciativa de oferecerem em conjunto, o Mestrado Interinstitucional.

As Faculdades Campo Real por ajustar os horários de aulas conforme as minhas necessidades.

A Fundação Araucária pelo auxílio financeiro através de bolsa.

A Diretoria e ao Departamento de Recursos Humanos do Grupo Superpão pela disponibilização dos dados.

A minha esposa Danielle e meus filhos Antonio Carlos e Anna Luisa, pela compreensão e entendimento nos momentos em que não pudemos estar juntos.

A minha mãe Marilena e meu irmão Antonio Carlos pelo apoio recebido.

A todos os colegas de curso pela amizade conquistada.

“Há homens que lutam um dia e são bons.
Há outros que lutam um ano e são melhores.
Há aqueles que lutam muitos anos e são muito bons.
Mas há os que lutam por toda a vida, esses são os imprescindíveis”.

Bertold Brecth

SUMÁRIO

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE SIGLAS

ACO	- Ant Colony Optimization
AD	- Árvores de Decisão
AF	- Análise Fatorial
Ags	- Algoritmos Genéticos
CE	- Computação Evolutiva
CLP	- Combinação Linear Padronizada
CN	- Computação Natural
DARPA	- Defense Advanced Research Projects
DM	- *Data Mining*
DW	- Data Warehouse
ECGA	- Eletrocardiograma Ambulatorial
FDL	- Função Discriminante Linear.
IA	- Inteligência Artificial
IC	- Inteligência Computacional
IDH	- Índice de Desenvolvimento Humano
KDD	- Knowledge Discovery Database
LM	- Levenberg-Marquadt
OLAP	- On-line Analytical Processing
RH	- Recursos Humanos
RNA	- Redes Neurais Artificiais
SAGRI	- Sistema Especialista Híbrido
WEKA	- Waikato Environment for Knowledge Analysis

RESUMO

O presente trabalho foi desenvolvido para predizer a satisfação 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, da empresa pesquisada, subsídios no momento de contratar seus colaboradores. Para tanto, utilizou-se de técnicas de predição em *Data Mining* (Mineração de Dados). Primeiramente desenvolveu-se um estudo do processo maior denominado *Knowledge Discovery in Database* (KDD), que em uma de suas fases abrange o *Data Mining*. Na seqüência elaborou-se um resumo das diversas técnicas de *Data Mining* e suas aplicações e entre elas, três foram selecionadas, por possuírem caráter dicotômico: Redes Neurais, Análise de Discriminante de Fisher e Regressão Logística. Aplicou-se na seqüência um questionário aos colaboradores da empresa com perguntas relacionadas à satisfação e características pessoais de cada indivíduo. As respostas desse questionário formaram o banco de dados numéricos que após sofrer algumas transformações e adaptações gerou um *Data Warehouse* (depósito de dados). Esses dados passaram então pelo primeiro conjunto de técnicas para sofrerem uma redução de dimensionalidade, através do Método das Componentes Principais e da Análise Fatorial e na seqüência foram analisados nas três técnicas de *Data Mining* selecionadas. Os resultados foram então testados e avaliados fazendo-se um comparativo entre as técnicas levando-se em consideração a margem de erro. No último estágio, um novo indivíduo foi avaliado utilizando os dados minerados no *Data Mining*. 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 da contratação de colaboradores da empresa.

Palavras Chaves: *Data Mining*, Análise Fatorial, Redes Neurais, Função Discriminante Linear de Fischer, Regressão Logística, Rotatividade de pessoal.

ABSTRACT

This research was developed to predict the satisfaction and the dissatisfaction of employees in a supermarket chain of retailers in order to reduce the employees' turnover, through information that makes possible to offer to the human resources department from the researched company, data at the moment of hiring its collaborators. For that, predictable techniques were used in *Data Mining*. First, it was developed a study of the biggest process, called Knowledge Discovery in Database (KDD), in which its phases encapsulates the *Data Mining*. Second, it was elaborated a sum-up of many techniques from *Data Mining* and its applications. Three of the techniques were chosen because they have dichotomy character: Neural Networks, Fisher Discriminator Analyses and Logistic Regression. In the sequence, a survey with the employees from the company was performed. There were questions related to the satisfaction and the individual characteristics. The answers of the survey were gathered and formed a numeric database that after some transformations and adaptation generated the Data Warehouse. These data were then submitted to the first set of techniques in order to be reduced dimensionally, through the Method of the Principal Components and Factor Analysis and after that, they were analyzed by the three techniques selected. The results were tested and evaluated, comparing the techniques taking into account the error spam. In the last stage, a new person was evaluated using the mined mining data from *Data Mining*. The conclusions identified that the Logistic Regression is the most suitable technique to the task and the results can be useful to improve the hiring process in the company.

KEYWORDS: *Data Mining*, Factor Analysis, Neural Networks, Fischer Discriminator Analysis, Logistic Regression, Personnel Turnover.

CAPÍTULO I

1. INTRODUÇÃO

Os primeiros anos do século XXI estão mostrando a dimensão da evolução tecnológica que o homem alcançou e que ainda almeja alcançar. Em todos os setores verifica-se um crescimento científico altamente qualificado. A Medicina caminha a passos largos juntamente com a Engenharia Genética na tentativa de descobrir terapias e tratamentos para doenças até então incuráveis, a Agricultura experimenta recordes de produtividade ano a ano, a Indústria Aeroespacial, encontra o décimo planeta no sistema solar e espera a descida do homem em Marte para dentro de 20 anos, a Meteorologia prevê condições climáticas com níveis de erro muito pequenos. Como tudo isso foi possível em tão pouco tempo? De que forma os saltos científicos tornaram-se cada vez mais amplos? Por que o homem demorou milhares de anos para dominar o fogo, descobrir a escrita, navegar, e, no entanto nos últimos 20 anos, o conhecimento humano teve um enriquecimento tão fabuloso? Todas essas perguntas são respondidas pensando-se que, vive-se a era da informação. O advento do computador, da telefonia móvel, dos satélites de comunicação, trouxe o mundo para dentro das empresas e das residências. A quantidade de informações e dados que se recebe por dia através de *e-mails*, *sites* de *Internet*, telefone e televisão é algo nunca visto. É esse *Big Bang*[1] de dados que permite que a tecnologia cresça de forma cada vez mais rápida.

Por outro lado, somente possuir dados não basta. Para construção do conhecimento científico necessita-se de ferramentas que auxiliem a leitura, seleção e a forma com que, os dados serão processados e convertidas em informações, ou seja, não adianta possuir quantidade e qualidade de dados e não saber o que fazer com eles.

A Estatística é a ciência dos dados e é ela que vem auxiliando o Homem a utilizar as informações com maestria. Usando-se recursos matemáticos e estatísticos criam-se algoritmos capazes de transformar dados abstratos em informações concretas e valiosas. No entanto a maior parte dos problemas estatísticos requer algum tempo para se chegar numa

---

[1] Information Explosion (Big Bang): denominação utilizada habitualmente para identificar a enorme quantidade de dados. Disponível no site: http://www.datawarehouse.inf.br/artigos.asp

solução. Até bem pouco tempo atrás, a análise de banco de dados estava limitada ao tempo de processamento e capacidade de solucionar equações complexas. Porém, a partir da década de 90, com a presença maciça do microcomputador nos bancos escolares, empresas e residências, novos *softwares* começaram a surgir, com possibilidades de armazenar e processar milhões de dados simultaneamente, tarefa que o cérebro humano não teria a mínima possibilidade de executar.

A análise de gigantescos bancos de dados através de algoritmos computacionais vem auxiliando atualmente governos, exércitos, universidades, empresas e pessoas na obtenção de valiosas informações que auxiliam significativamente na tomada de decisão, minimizando erros, custos e tempo e maximizando acertos, lucros, rapidez e satisfação.

Existem vários técnicas de análise de dados, a maioria delas derivadas da Estatística Clássica, pois utilizam conceitos de distribuição normal, variância, desvios, intervalos de confiança, análises de discriminante e principalmente trabalham com resultados. Entre o aglomerado de técnicas existente, um conjunto delas vem sendo largamente utilizado, trata-se do *Data Mining* ou mineração de dados.

Data Mining é um conjunto de técnicas, sujeitas a erros, que permite buscar em um grande banco de dados, informações que aparentemente, estão camufladas, permitindo, com isso, agilidade nas tomadas de decisão. *Data Mining* surgiu com o armazenamento maciço de dados e a necessidade de transforma-los em informação.

Quando determinados padrões de comportamento, começam a se repetir, com freqüência em determinados grupos sociais, clientes ou funcionários, as ferramentas de *Data Mining* indicam a presença de *insights* dentro desses grupos.

Um exemplo clássico de *Data Mining* foi desenvolvido pela *Wall-Mart*[2], empresa varejista norte-americana. A empresa descobriu que o perfil do consumidor de cervejas era semelhante ao de fraldas. Eram homens casados, entre 25 e 30 anos, que compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa. Com base na verificação destas hipóteses a *Wall-Mart* optou por colocar fraldas ao lado das cervejas nas gôndolas de suas lojas. Como resultado o consumo cresceu 30% às sextas-feiras baseado na conexão de hipóteses desenvolvidas pelo *Data Mining*.

Empresas varejistas do ramo de supermercados trabalham com o mesmo ramo de negócios da *Wall-Mart*. Guardadas as devidas proporções com relação ao volume de

faturamento, número de lojas e número de funcionários, os problemas de ambas são semelhantes, pois vendem a varejo produtos em supermercados. Assim como a *Wall-Mart,* os diretores e colaboradores dessas empresas almejam condições de trabalho, níveis salariais e lucratividades maiores. Todo conhecimento novo que pode ser empregado para reverter em beneficio da empresa e de seus colaboradores é bem vindo.

Um dos maiores problemas enfrentados atualmente dentro destas empresas é o que diz respeito à alta rotatividade de pessoal. Desta forma o *Data Mining* 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 subsidie o departamento de recursos humanos com informações importantes no momento do recrutamento e seleção do pessoal contratado, minimizando a alta rotatividade.

1.1 OBJETIVOS DO TRABALHO

O objetivo do presente trabalho 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 pré-dizer 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.

1.2 ESTRUTURA DO TRABALHO

O presente trabalho foi dividido em seis capítulos e seus conteúdos estão descritos na seqüência:

---

[2] Disponível no site: http://www.datawarehouse.inf.br/artigos.asp

- Capítulo 1: Esse capítulo é formado pela introdução ao assunto, a descrição do problema e os objetivos a serem alcançados.

- Capítulo 2: Esse capítulo aborda conceitos fundamentais de sistemas, empresa como sistema e o sistema de recursos humanos, descreve o conceito e as fases do KDD (*Knowledge Discovery in Databases*) e do *Data Mining*, além de elencar dez técnicas de *Data Mining*. Ainda contém a revisão bibliografia e cita os ramos do conhecimento onde o *Data Mining* vem sendo utilizado e também contém trabalhos de outros autores em que foram utilizadas técnicas de *Data Mining*
- Capítulo 3: Descreve toda a fundamentação matemática envolvida na solução do problema incluindo os modelos matemáticos da técnica de redução de dimensionalidade e das três técnicas de predição.
- Capítulo 4: Descreve a aplicação das técnicas ao problema, a metodologia empregada e os resultados obtidos.
- Capítulo 5: Apresenta as conclusões e os tópicos principais que deixam margem para futuros trabalhos relacionados na área.

1.3 DESCRIÇÃO DO PROBLEMA.

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, *drugstore,* etc, para realizar compras. No Brasil, volume de capital envolvido é algo fabuloso e beira R$300.000.000,00 [3] por dia. 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.

[3] Na realidade a cifra de R$300.000.000,00 por dia, corresponde as cinco maiores empresas brasileiras de acordo com dados da APRAS – Associação Brasileira de Supermercados no ano de 2004.

> "Por muitos anos se pensou que o obstáculo que segura o desenvolvimento de uma empresa fosse o capital. Era uma crença generalizada. Todavia, é a inabilidade de uma empresa em recrutar e manter uma boa força de trabalho que constitui o principal obstáculo para a produção. Não existe nenhum projeto baseado em boas idéias, vigor e entusiasmo que tenha sido interrompido por falta de caixa ou recursos financeiros. Existem empresas que cresceram e cujo crescimento foi parcialmente bloqueado ou dificultado porque não puderam manter uma força de trabalho eficiente e entusiasmada" (CHIAVENATO, 1999).

Num mundo informatizado, receber uma saudação na entrada e saída do supermercado, tornar-se conhecido do operador de caixa que o trata pelo seu próprio nome e ser atendido pela mesma pessoa no açougue, feira e panificadora é desejo de qualquer cliente.

Esse modelo personalizado de atender os clientes é uma constante nas lojas de supermercados pesquisadas, pois possuem característica interiorana, personalizando o atendimento a cada cliente. No entanto conseguir conscientizar os funcionários deste modelo de atendimento e sobre tudo manter o pessoal motivado para praticá-lo não é tarefa fácil, principalmente quando o grupo se quebra com conseqüente saída e substituição de colaboradores.

Segundo CHIAVENATO (1999), para haver produtividade no trabalho é necessário que "os empregados sintam que o trabalho é adequado às suas capacidades e que estão sendo tratados eqüitativamente. As pessoas despendem a maior parte de suas vidas no trabalho e isto requer uma identidade com o trabalho que fazem. Empregados satisfeitos não são necessariamente os mais produtivos, mas empregados insatisfeitos tendem a se desligar da empresa, se ausentar freqüentemente e produzir pior qualidade do que empregados satisfeitos. A felicidade na organização e a satisfação no trabalho são fortes determinantes do sucesso organizacional".

Os problemas gerados pela rotatividade de pessoal são muitos, mas pode-se destacar dois como mais importantes:

- A troca rotineira de funcionários gera custos de treinamento, rescisões contratuais e maior índice de ações trabalhistas.
- A rotatividade afeta significativamente o cliente das lojas, que se sente bem, quando é recebido e atendido por funcionários já conhecidos. Perde-se com a rotatividade o lado social e humano, característica marcante das lojas estudadas.

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?

A garimpagem de dados realizada ao longo deste trabalho tentará responder as perguntas acima mencionadas.

CAPÍTULO II

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta a revisão bibliográfica, contento as técnicas de *Data Mining* sendo utilizadas na resolução de problemas em diversas áreas e também contém pequenos resumos de trabalhos acadêmicos que apresentam aplicação de técnicas em *Data Mining*.

Antes de iniciar a descrição do processo de *Data Mining*, algumas considerações e conceitos com relação à empresa, sistemas e recursos humanos devem ser colocados, visto que o trabalho se refere às interligações entre indivíduos e seus ambientes de trabalho.

2.1 O CONCEITO TRADICIONAL DE EMPRESA

Existem inúmeros conceitos de empresa na literatura atual, mas entre tantos, um deles merece destaque:

> "A empresa constitui o ambiente dentro do qual as pessoas trabalham e vivem a maior parte de suas vidas. Nesse contexto as pessoas dão algo de si mesmas e esperam algo em troca, seja a curto ou em longo prazo A maneira pela qual esse ambiente é moldado e estruturado influencia poderosamente a qualidade de vida das pessoas. Mais do que isso: influencia o próprio comportamento e os objetivos pessoais de cada ser humano. E isto, conseqüentemente, afeta o próprio funcionamento da empresa" (CHIAVENATTO, 1994).

2.2 A EMPRESA COMO SISTEMA

Toda empresa tem a sua própria cultura organizacional, representada, entre outros, por fatores como: a filosofia administrativa, políticas de atuação no mercado, tradição e imagem e processos. Esses entre outros fatores organizacionais devem funcionar, na medida do possível, de forma organizada e dinâmica. Desse modo, a empresa também deve ser considerada como um sistema.

O objetivo central de uma empresa como sistema é a transformação de bens, de valores, de recursos humanos, materiais e financeiros.

2.3 ORIGEM DOS SISTEMAS

O filósofo inglês Herbert Spencer (1820-1904) afirmava, que um organismo social assemelha-se a um organismo individual nos seguintes traços essenciais:

- no crescimento;
- no fato de tornar-se mais complexo à medida que cresce;
- no fato de que se tornando mais complexo, suas partes exigem uma crescente interdependência mútua; e
- por que em ambos os casos há crescente integração acompanhada por crescente heterogeneidade.

No início da década de 30, o filósofo e cientista social belga Lévi-Strauss, dizia que "uma estrutura oferece um caráter de sistema, consistindo em elementos combinados de tal forma que qualquer modificação de um deles implica na modificação de todos os outros".

2.4 A EXPERIÊNCIA DE BERTALANFFY

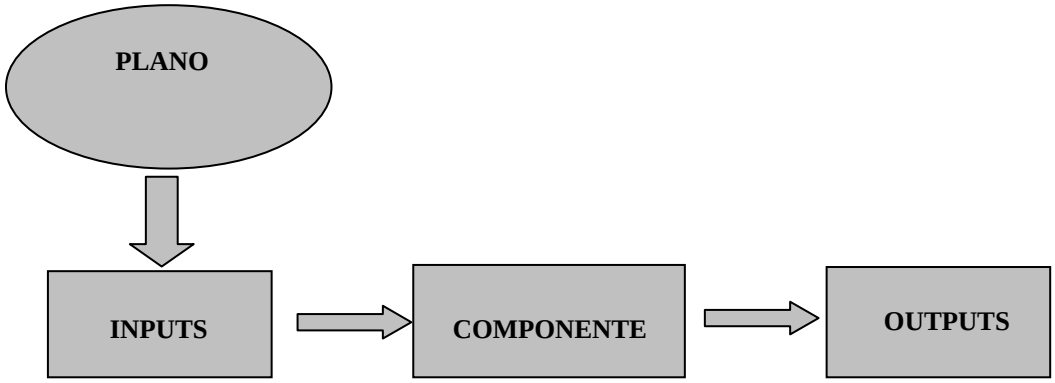
O vocábulo "sistema", tal como hoje é concebido na área administrativa, está relacionado com as experiências do biólogo alemão Ludwig Von Bertalanffy. Esse cientista na década de 50, pesquisando comportamento de organismos vivos, constatou que, independentemente de sua forma e de suas características, os seres biológicos possuem vários pontos em comum.

Deve-se a Bertalanffy a divulgação de expressões como *feedback* (realimentação, retroação); *input* (entrada); *output* (saída), expressões estas incorporadas à cibernética, eletrônica e computação.

Bertalanffy estendeu seus estudos a outros tipos de organismos: sociais, mecânicos, eletrônicos, etc; confirmando que tal como acontece com os seres vivos, esses organismos não naturais conservam igualmente, certas características comuns. Entre essas características a principal é o Objetivo Atingido.

Com base nas descrições de Bertalanffy pode-se definir um sistema como sendo um conjunto de componentes interligados e capazes de transformar uma série de *inputs* numa série de *outputs* para atingir objetivos dentro de um plano pré-estabelecido. A figura 2.1 ilustra melhor a definição.

FIGURA 2.1 – DEFINIÇÃO DE SISTEMA SEGUNDO BERTALANFFY



FONTE: AUTOR

2.5 O SISTEMA DE RECURSOS HUMANOS

Os sistemas empresariais podem ser abertos ou fechados. Sistemas fechados são aqueles em que os processos e produtos estão protegidos por monopólios ou patentes. Exemplos típicos de sistemas fechados são os monopólios, que realizam suas transações limitadamente em relação ao ambiente.

Por outro lado, os sistemas abertos trocam matéria e energia com o ambiente regularmente.

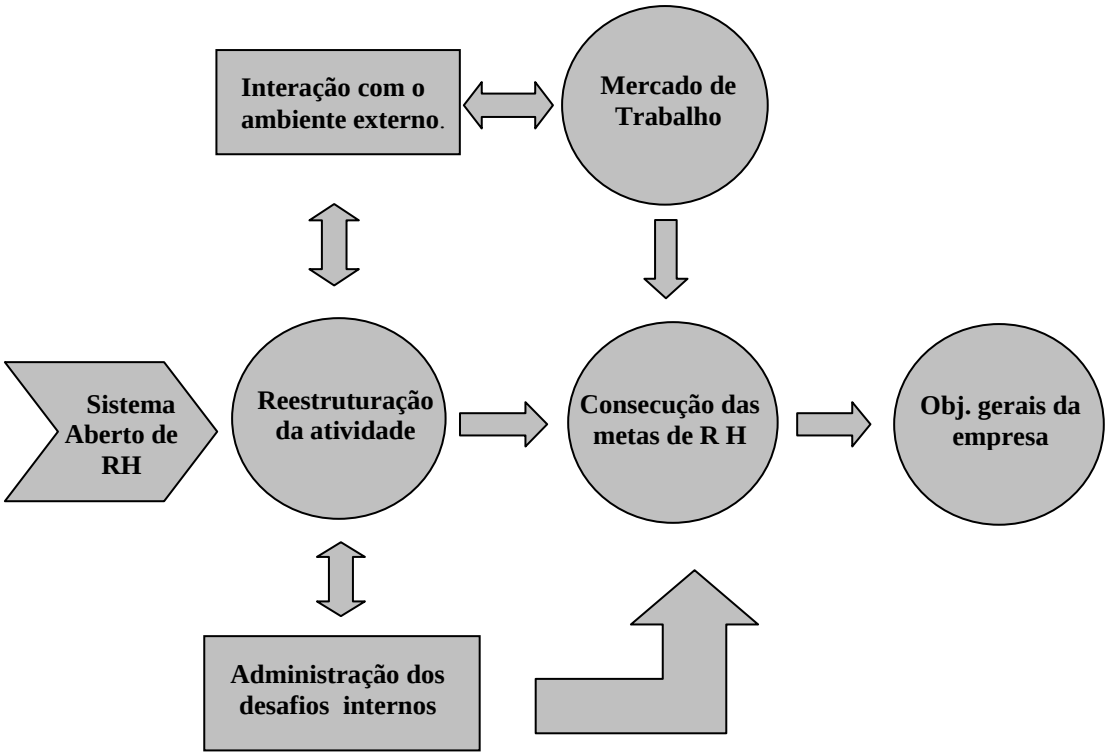
Na empresas os sistemas abertos recebem o nome de dinâmicos, pois estão em profunda interação, exportando e importando produtos, serviços e conhecimento, entre os diversos sistemas: financeiro, compras, marketing, recursos humanos, etc.

A Atividade de Recursos Humanos somente pode se manifestar plenamente em uma empresa em um sistema administrativo aberto.

O esquema da figura 2.2 retrata os dois ramos da função RH, ou seja, a ramificação externa de relações da empresa com o ambiente onde está inserida, com seus problemas de mercado, econômico e social da própria nação.

Enquanto que na ramificação externa as funções concentram-se na resolução dos problemas e desafios da própria empresa na gestão eficiente dos colaboradores (mão-de-obra) disponíveis.

FIGURA 2.2 – RAMIFICAÇÕES DA FUNÇÃO DE R.H. COM O AMBIENTE.



FONTE: AUTOR

Tanto os ramos externos como internos devem auxiliar de maneira decisiva, no sentido de atingir as metas e objetivos propostos no planejamento de RH.

2.6 A NOVA FUNÇÃO DO R.H NA EMPRESA MODERNA

Os departamentos de RH. sempre se preocuparam 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.

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:

- Ambiente Organizacional, aí situados os desafios internos – estudos científicos das relações de trabalho – e externos - cultura ambiental, influência do mercado consumidor etc.

- Objetivos da organização: revisão de prioridades e metas da empresa.
- Estrutura de cargos, responsabilidades e níveis de comando.
- Motivação e liderança de equipes de trabalho.
- Relação de poder.

A identificação desses itens é preponderante para que o sistema de RH seja interpretado de forma moderna, dinâmica e arrojada.

2.7 POLÍTICAS DE RECURSOS HUMANOS

As políticas de RH – cargos e salários, treinamento, avaliação, carreira etc. – estão subordinadas à filosofia empresarial e devem possuir flexibilidade, adaptando-se aos objetivos organizacionais.

Se por um lado à filosofia da empresa é algo duradouro e estável, por outro as políticas de RH são mutáveis e dinâmicas e dependem de fatores externos como:

a) das reações do mercado;
b) da influência do Estado;
c) 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 prazo. É na implantação dessas metas que se deve 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.

2.8 OBJETIVO DO SISTEMA DE RH.

Os objetivos do sistema aberto de Recursos Humanos podem ser classificados da seguinte maneira:

2.8.1 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.

2.8.2 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.

2.8.3 Objetivos Funcionais

Manter nem nível adequado seus procedimentos em função das necessidades efetivas de mão-de-obra treinada, consciente e responsável.

2.8.4 Objetivos Individuais

Oferecer assistência aos funcionários na consecução de suas metas individuais, na medida que a administração participativa tende a se expandir na organização.

2.9 O KDD E O DATA MINING

É praticamente impossível falar em *Data Mining* sem antes falar no KDD. Verdadeiramente o *Data Mining* é uma parte de um processo muito maior denominado KDD, do inglês *Knowledge Discovery in Databases*, ou seja, a Busca de Conhecimentos em Banco de Dados, sendo o *Data Mining* apenas uma das fases desse processo.

O KDD "consiste em um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, em conjunto de dados" (Fayyad, 1996).

Examinando os termos contidos nessa definição com mais detalhes, tem-se que:

Dados: correspondem ao conjunto de fatos F, por exemplo, casos em um banco de dados.

Padrão: é uma expressão E em uma linguagem L, descrevendo fatos em um

subconjunto F_c de F.

Processo: o processo em KDD é composto por várias etapas, que envolve preparação de dados, busca de padrões, avaliação do conhecimento e refinamento envolvendo iteração depois de modificação. O processo é assumido como não trivial por ter algum nível de busca autônoma. Por exemplo, o cálculo da média das notas de alunos em uma sala de aula não é qualificado como um processo KDD, por ser uma tarefa trivial.

Validade: A descoberta de padrões deve ser válida sobre novos dados com determinado grau de certeza.

Novo: os padrões são novos podendo ser a comparação de valores correntes com prévios já esperados ou ainda de que forma o conhecimento novo está relacionado com um antigo.

Potencialmente útil: os padrões devem levar a uma ação útil, sendo medida por alguma função de utilidade.

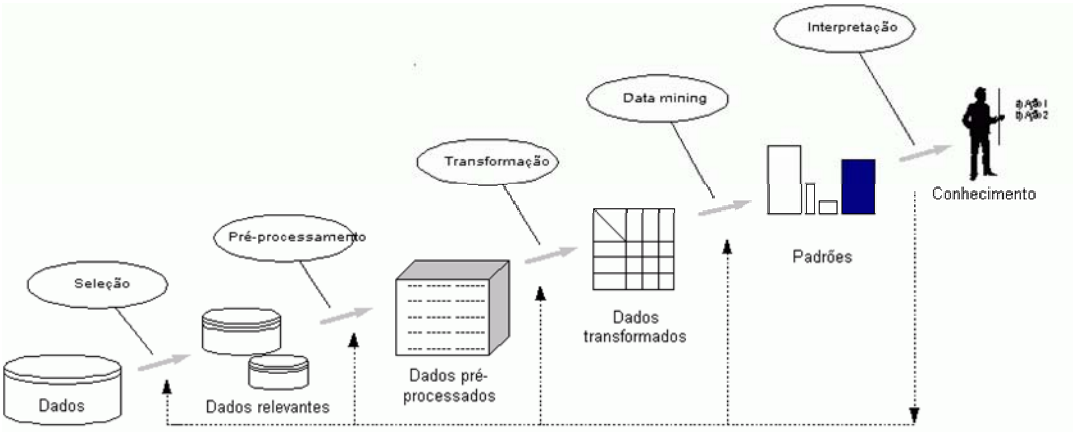
Compreensível: um objetivo do KDD é gerar conhecimento compreensível por seres humanos para facilitar a compreensão dos dados subjacentes. Se o conhecimento não é compreensível pelo usuário ele não é capaz de interpretá-lo ou validá-lo, portanto não podendo utilizar esse conhecimento em tomadas de decisões.

A noção de conhecimento é bastante subjetiva e depende muito das noções do usuário. Além do mais o que pode ser conhecimento para um determinado usuário, pode não ser para outro.

2.10 AS FASES DO PROCESSO KDD.

Para melhor entendimento das fases do processo KDD a utilização da figura 2.3 é esclarecedora, pois apresenta uma visão clara dos passos, onde o *Data Mining* é apenas uma das fases que ocorre sobre os dados que foram transformados nos passos anteriores.

FIGURA 2.3 – FASES DO PROCESSO KDD



FONTE: FAYAD-1996

Segue uma pequena descrição de cada um dos seis passos do processo KDD (Fayyad, 1996).

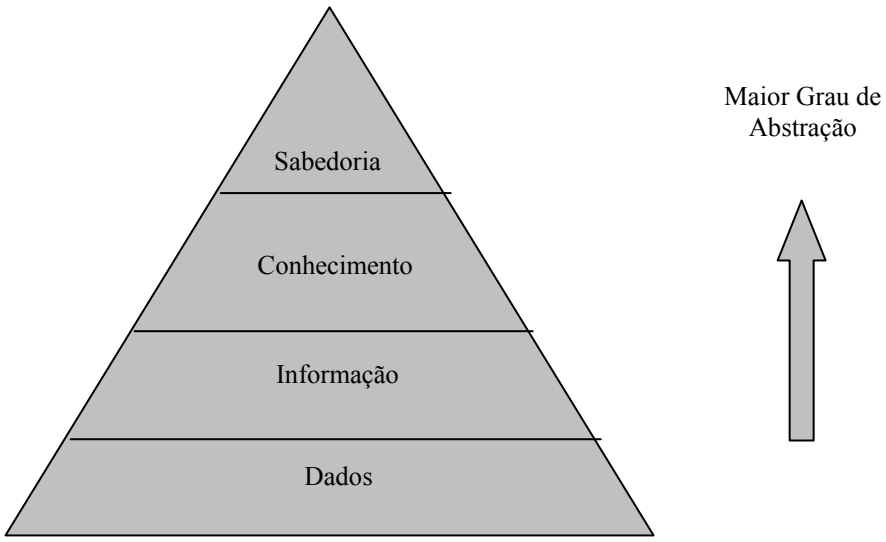
1. O primeiro passo para realizar o processo KDD consiste em obter a compreensão do domínio da aplicação e as metas do usuário final. É preciso observar quais as bases de dados disponíveis e o volume de dados associados de alguma forma com a meta estabelecida.
2. Em seguida deve-se criar um conjunto de dados meta, selecionando um conjunto de dados, ou estabelecendo um subconjunto de variáveis ou exemplos de dados, sobre os quais a descoberta será executada.
3. Alguns fatores colaboram com os erros de reconhecimento de padrões. O ambiente de extração das características freqüentemente apresenta ruídos, distorções, etc. Para diminuí-los deve-se realizar operações básicas, tais como remover os ruídos ou valores incoerentes, coletar as informações necessárias para o modelo e decidir sobre estratégias para controlar campos de dados perdidos.
4. Efetuar a redução e projeção dos dados, encontrando características úteis para representá-los, dependendo do objetivo da tarefa. Deve-se também usar a redução de dimensionalidade ou método de transformação para reduzir o número efetivo de variáveis em consideração.
5. Na fase de *Data Mining*, primeiramente deve-se definir a tarefa de mineração que será feita, ou seja, classificação, regressão, agrupamento, etc. Uma vez definida a tarefa, deverá ser eleito o algoritmo a ser usado para encontrar os

padrões nos dados.A mineração dos dados é executada então, por meio do algoritmo que, para apresentar os padrões encontrados, adotará uma forma representacional particular, como regras de classificação, árvores, gráficos de agrupamentos, etc.

6. Nesse passo é importante interpretar os padrões minerados, possivelmente retornando a qualquer um dos passos anteriores para novas iterações, caso necessário.
7. Finalmente é necessário consolidar o conhecimento descoberto, incorporando-o ao sistema global, ou simplesmente documentando-o e relatando-o para as partes interessadas. Isso também inclui checar e resolver possíveis conflitos com conhecimentos previamente extraídos ou conhecidos.

Sob um outro ponto de vista o KDD pode ser representado através de estágios que compõem uma pirâmide (figura 2.4). Na medida que os níveis da pirâmide vão subindo o refinamento dos dados, oferecendo abstração cada vez mais intensa, passando pelos níveis de informação, conhecimento e culminando com a sabedoria, último nível do KDD.

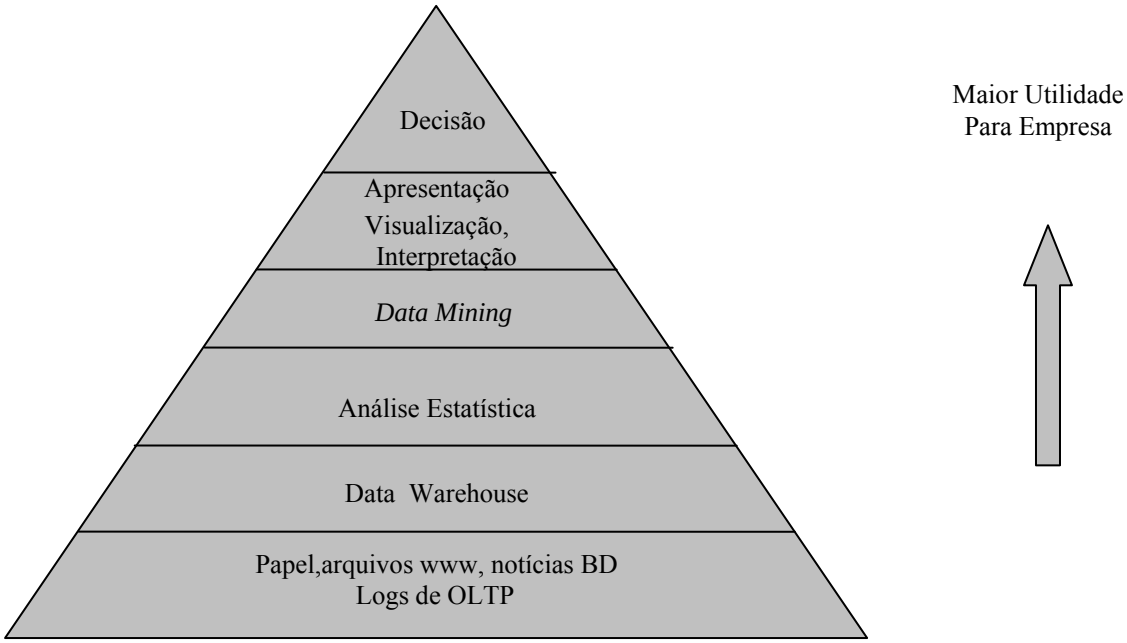
FIGURA 2.4 - PIRÂMIDE DOS ESTÁGIOS DO PROCESSO KDD



FONTE: ANAIS DA XII ESCOLA REGIONAL DE INFORMÁTICA DA SBC – PARANÁ

O Diagrama da pirâmide, quando adaptado a uma empresa moderna adquire o formato da figura 2.5 (Navega, 2002). Os níveis recebem uma nomenclatura que possibilite a leitura pelo administrador.

FIGURA 2.5 – PIRÂMIDE ADAPTADA PARA A EMPRESA MODERNA



FONTE: ANAIS DA XII ESCOLA REGIONAL DE INFORMÁTICA DA SBC – PARANÁ

Todos os passos do KDD são importantíssimos para a evolução do processo, porém o foco deste trabalho concentra-se em duas fases: o Banco de Dados e o processo de *Data Mining* e suas técnicas.

2.11 O BANCO DE DADOS

Qualquer trabalho de pesquisa que se deseja fazer começa necessariamente tendo pela frente um banco de dados, porém quando se fala em Banco de Dados muitas perguntas podem surgir e algumas se encontram relacionadas abaixo:

- De que tamanho deve ser um Banco de Dados?
- Quais informações devem conter?
- Pode-se usar dados existentes ou existe necessidade de novos dados?
- De que forma os dados devem ser apresentados?
- Os dados são relevantes para o resultado desejado?
- Pode-se agrupar os dados em subconjuntos menores?
- Os dados são confiáveis?

Essas e tantas outras perguntas somente são respondidas à medida que o

pesquisador inicia o processo de conhecimento criterioso do problema que possui em mãos. O aprofundamento no problema em si gera sabedoria e discernimento para escolha dos dados potencialmente úteis e que serão relevantes na busca de soluções. Portanto não existem regras claras e definidas na escolha do Banco de Dados, cada problema é único e deve ser tratado como tal.

2.12 TABELA DE DADOS

Geralmente quando se trabalha com *Data Mining*, os dados estão alocados em tabelas contendo linhas e colunas em formato matricial. Uma característica importante em uma tabela de dados é definida como sendo um conjunto de linhas que compartilham o mesmo valor ou o mesmo significado em duas ou mais colunas. É através da análise destas interações, entre linhas e colunas, que as técnicas em *Data Mining*, descobrem novos conhecimentos. Desta forma a necessidade do Banco de Dados estar no formato matricial, é imprescindível para utilização do *Data Mining*.

2.13 DATA WAREHOUSE

> "Data Warehouse, ou depósito de dados, é um sistema de gerenciamento de banco de dados relacional, desenvolvido especificamente para as necessidades de sistemas de processamento de transações e por esta razão tem uma associação muito forte com *Data Mining*. As possibilidades de mineração de dados podem ser intensificadas se os dados estiverem armazenados em um *Data Warehouse"* (LOUZADA e DINIZ, 2002).

Segundo INMON (1993), *Data Warehouse* "é uma coleção de dados com quatro características: tópico-orientado, integrado, tempo-variante e não volátil".

a) Tópico-orientado: os dados são definidos e organizados em assuntos de negócios, em vez de aplicações. Por exemplo: uma companhia de seguro organiza seus dados de seguro por consumidor, prêmios e sinistros e não utilizando o produto: seguro de automóveis, de vida, de incêndio.

b) Integrado: quando os dados residem em várias aplicações separados no ambiente operacional é provável que exista uma codificação inconsistentes nos mesmos. Por exemplo, em uma aplicação o sexo do indivíduo pode ser codificado como "f" ou "m", ou ainda "0" ou "1". Quando os dados são

transferidos de ambiente operacional para o *Data Warehouse*, eles assumem um código convencional consistente.

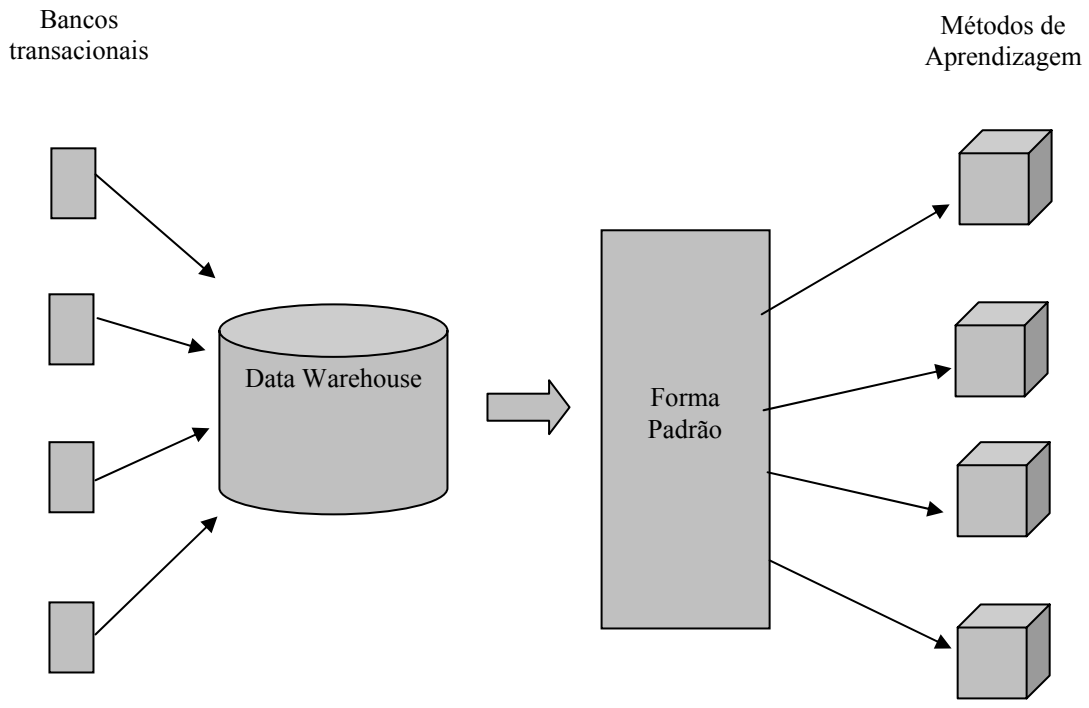
c)Tempo variante: o *Data Warehouse* contem um espaço para armazenar dados antigos de 5 a 10 anos, por exemplo, que podem ser usados em comparações, tendências e previsões, sendo que esses dados não são atualizados.

d) Não Volátil: uma vez que os dados entram no *Data Warehouse* não são atualizados ou mudados, são somente carregados ou acessados.

2.14 DATA WAREHOUSE PARA *DATA MINING*

Um modelo de armazenamento de dados para *Data Mining* é ilustrado na figura 2.6. Geralmente, bancos de dados em grandes organizações, estão espalhados entre departamentos, setores e lojas. Desta forma um *Data Warehouse* deve ser construído, para aglutinar esses dados de forma centralizada, disponibilizando-os para futuros processos analítico e decisórios.

FIGURA 2.6 – MODELO DE ARMAZENAGEM DE DADOS.



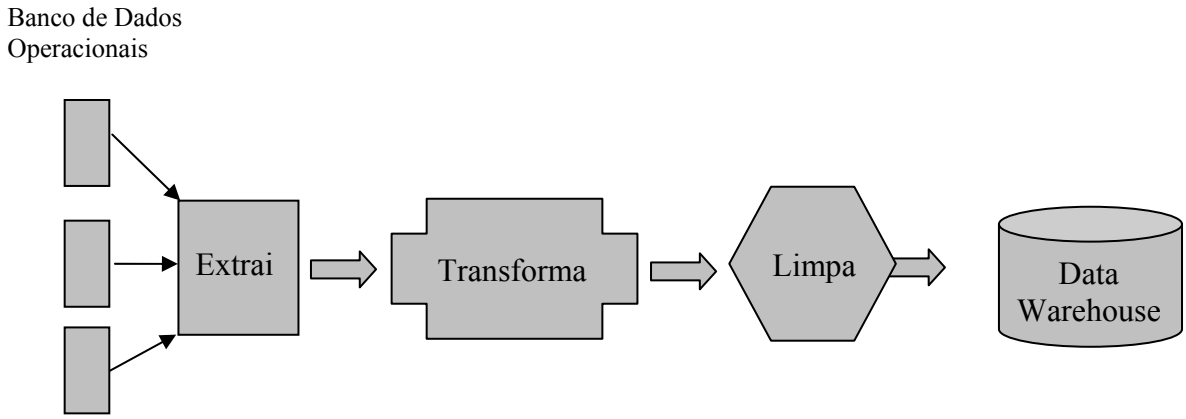
FONTE: LOUZADA E DINIZ

Os dados provenientes do banco de dados não estão preparados necessariamente para serem minerados. Existe uma longa caminhada, até que os dados transacionais sejam transformados e armazenados em um *Data Warehouse*, para daí sim, num formato padrão,

possa produzir métodos de aprendizagem.

A figura 2.7 ilustra as tarefas necessárias a serem realizadas para que os dados originais, alocados em um banco de dados operacionais, sejam formatados para padrões desejáveis e a partir desde momento armazená-los no *Data Warehouse*.

FIGURA 2.7 – TAREFAS NECESSÁRIAS PARA FORMAÇÃO DE UM DATA WAREHOUSE



FONTE: LOUZADA E DINIZ

As tarefas seguem abaixo:

a) Extração: dados são extraídos de diferentes fontes em diversos formatos

b)Transformação: dados brutos são transformados em dados mais qualificados para apoio à decisão

c) Limpeza: os campos de dados são verificados procurando-se por inconsistências ou por valores faltantes. Registros errados são solucionados ou eliminados.

d) Integração: dados de múltiplos bancos de dados e outras fontes são integrados em um *Data Warehouse* central.

Somente a existência de um *Data Warehouse* não resolve todos os problemas pertinentes a preparação de dados. Muitas vezes são necessárias transformações adicionais para compatibilizar com as aplicações a serem desenvolvidas no *Data Mining*.

Um problema enfrentado quando se trabalha com grandes bancos de dados diz respeito ao tempo de resposta necessário para processamento das informações. Para minimizar esse problema pode-se lançar mão de um OLAP – *On line Analytical Processing*, que é uma extensão do *Data Warehouse* e é uma ferramenta que integra, *on line* a nova entrada de dados com o processamento das variáveis a análise e proporciona respostas rápidas (DILLY, 1998 e INMON, 1996).

2.15 *DATA MINING* - CONCEITOS

A literatura atual oferece várias definições para *Data Mining*, cada qual com suas propriedades e suas características, algumas até mesmo conflitantes, porém, a maioria das definições afunila sempre para um conceito único: *Data Mining* é uma das fases do processo de KDD, onde dados são minerados através de algoritmos computacionais, objetivando produzir novas informações.

Seguem algumas definições:

"*Data Mining* é um passo do processo, consistindo de algoritmos de mineração que, sob algumas limitações aceitáveis de eficiência computacional, produz uma enumeração de padrões E_j sobre F" (FAYYAD, 1996).

"*Data Mining* é o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano" (CARVALHO, 2002).

"*Data Mining* é uma ferramenta utilizada para descobrir novas correlações, padrões e tendências entre as informações de uma empresa, através da análise de grandes quantidades de dados armazenados em *Data Warehouse* usando técnicas de reconhecimento de padrões, Estatísticas e Matemáticas" (NIMER & SPANDRI, 1998).

"*Data Mining* é um processo de extração não trivial de conhecimentos implícitos previamente conhecidos e potencialmente úteis em dados de um banco de dado" (SHAPIRO, 1991).

"*Data Mining* é a busca por relações e características globais que estão" escondidas "em uma vasta quantidade de dados, tal como a relação entre dados de pacientes e seus diagnósticos médicos. Estas relações apresentam valiosos conhecimentos a respeito do banco de dados" (HOLSHEMIER E SIEBES, 1994).

"*Data Mining* refere-se ao" uso de uma variedade de técnicas para identificar informações úteis em bancos de dados e a extração dessas informações de tal maneira que elas possam ser usadas em áreas tais como teoria de decisão, estimação, predição e previsão.Os bancos de dados são geralmente volumosos,

> e na forma que se encontram nenhum uso direto pode ser feito deles; as informações escondidas nos dados é que são realmente úteis ”(GUIDE, 2000).

> A IBM define *Data Mining* como “um processo de extração de informações previamente desconhecidas, válidas e com capacidade de proporcionar ações, provenientes de uma grande base de dados e então estudá-las na tomada de decisões de negócios cruciais".

2.16 RAMOS DO CONHECIMENTO QUE ENVOLVE O *DATA MINING*.

Apesar de ser um mecanismo bastante novo, o *Data Mining* possui origens mais remotas e usa como alicerce linhagens da ciência que o sustentam formando um tripé.

Segundo Andreatto (2002) “o *Data Mining* é um campo interdisciplinar que Envolve a Estatística, a Inteligência Artificial e o Aprendizado de Máquina”.

2.16.1 Estatística

O *Data Mining* possui a maior parte de sua estrutura herdada da estatística clássica, sem a qual seria impossível falarmos em *Data Mining*.

O uso de conceitos estatísticos tais como, distribuição normal, variância, análise de regressão, desvios simples, análises de conjuntos, análises de discriminantes e intervalos de confiança são utilizados largamente para analisar dados relacionados entre si. As mais avançadas análises estatísticas utilizam esses conceitos e o *Data Mining* também.

2.16.2 Inteligência Artificial.

A Inteligência Artificial, ou IA, é construída a partir dos fundamentos da heurística, em oposto à estatística, tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos. Em função desse *approach*, ela requer um impressionante poder de processamento, que era impraticável até os anos 80, quando computadores começaram a oferecer um bom poder de processamento e a preços mais acessíveis.

A credibilidade da aplicação efetiva de soluções baseadas em IA ficou comprometida em função às inúmeras discussões teóricas sobre a significância e magnitude do termo IA, envolvendo inclusive aspectos filosóficos, no sentido até onde se pode criar uma inteligência artificial.

Desta forma surgiu então o termo Inteligência Computacional, ou, IC, como sendo uma área da computação concentrada na implantação de solução colaborativas e não competitivas, no sentido em que visa implementar em sistemas complexos do mundo real, soluções computacionais com algumas características de comportamento inteligente para contribuir na busca por resultados melhores e mais compreensíveis.

Segundo Palazzo (2003) a área denominada Inteligência Computacional, (IC) é formada pelo estudo de sistemas *Fuzzy*, Redes Neurais e Computação Evolutiva (CE) que compreende uma ramificação da Ciência da Computação. A Inteligência Artificial (IA), por sua vez, em conjunto com outras áreas, tais como Vida Artificial, Geometria Fractal, Teoria do Caos, Sistemas Complexos, que delimitam um campo conhecido como Computação Natural (CN).

É importante salientar que enquanto a IA ocupa-se de aspectos teóricos e conceituais da representação de modelos de inteligência e de conhecimento, a IC ocupa-se, de modo mais prático, da implementação computacional desses modelos na busca de solução e otimização de problemas complexos.

2.16.3 Aprendizado de Máquina.

É a terceira e última linhagem do DM é chamada *machine learning*, que pode melhor ser descrita como o casamento da estatística com a Inteligência Artificial. *A machine learning* tenta fazer com que os programas de computador "aprendam" com os dados que eles estudam, tal que esses programas tomem decisões diferentes baseadas nas características dos dados estudados, usando a estatística para os conceitos fundamentais, e adicionando mais heurísticas avançadas da Inteligência Artificial e algoritmos para alcançar os seus objetivos.

De muitas formas, o DM é fundamentalmente a adaptação das técnicas da *machine learning* para aplicações de negócios. Desse modo, pode-se descreve-lo como a união dos históricos e dos recentes desenvolvimentos em estatística, em IA e *machine learning*. Essas técnicas são usadas juntas para estudar os dados e achar tendências e padrões nos mesmos. Hoje, o DM tem experimentado uma crescente aceitação nas ciências e nos negócios que precisam analisar grandes volumes de dados e achar tendências que eles não poderiam achar de outra forma.

2.17 TAREFAS EM *DATA MINING*

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:

Associação: estuda um padrão de relacionamento entre itens de dados. Por exemplo, uma análise das transações de compra em um supermercado pode encontrar itens que tendem a ocorrerem juntos em uma mesma compra (fraldas e cervejas), por exemplo. Os resultados desta análise podem ser úteis na elaboração de catálogos e *layout* de prateleiras de modo que produtos a serem adquiridos na mesma compra fiquem próximos um do outro. Essa tarefa é considerada descritiva, pois busca identificar padrões em dados históricos.

Classificação: consiste em examinar as características de um objeto ou situação e atribuir a ele uma classe pré-definida. Ou seja, esta tarefa objetiva a construção de modelos que permitam agrupamento de dados em classes. Essa tarefa é considerada preditiva, pois uma vez que as classes são definidas, ela pode prever automaticamente a classe de um novo dado. Por exemplo, uma população pode ser dividida em categorias para avaliação de concessão de crédito com base em um histórico de transações de créditos anteriores. Em seguida, uma nova pessoa pode ser enquadrada, automaticamente, em uma categoria de crédito específica, de acordo com suas características. Existem várias técnicas para a classificação: Árvores de Decisão, Regressão Logística, Redes Neurais e Algoritmo Genético.

Estimativa (regressão): objetiva definir um valor numérico de alguma variável desconhecida a partir dos valores de variáveis conhecidas. Exemplos de aplicação são: estimar a probabilidade de um paciente sobreviver dado o resultado de um conjunto de diagnósticos de exames; predizer quantos carros passam em determinado pedágio, tendo alguns exemplos contendo informações como: cidades mais próximas, preço do pedágio, dia da semana, rodovia em que está localizado o pedágio, entre outros. Essa tarefa é considerada preditiva. As

técnicas utilizadas nessa tarefa são: Redes Neurais, Regressão Linear, Análise de Discriminante.

d) Clusterização (análise de agrupamentos-*cluster analysis*): as informações podem ser particionadas em classes de elementos similares. Neste caso, nada é informado ao sistema a respeito das classes existentes. O próprio algoritmo descobre as classes a partir das alternativas encontradas na base de dados, agrupando assim um conjunto de objetos em classes de objetos semelhantes. Por exemplo, uma população inteira de dados sobre tratamentos de uma doença pode ser dividida em grupos baseados na semelhança de efeitos colaterais produzidos; acessos a *Web* realizados por um conjunto de usuários em relação a um conjunto de documentos podem ser analisados para revelar clusters ou categorias de usuários. Essa tarefa é considerada descritiva. A técnicas utilizadas nessa tarefa é chamada de Análise de Cluster e utiliza uma área da Estatística denominada Análise Multivariada.

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.

Seguem as técnicas de *Data Mining* pesquisadas neste trabalho e um pequeno resumo de cada uma delas.

2.18 TÉCNICAS EM *DATA MINING*

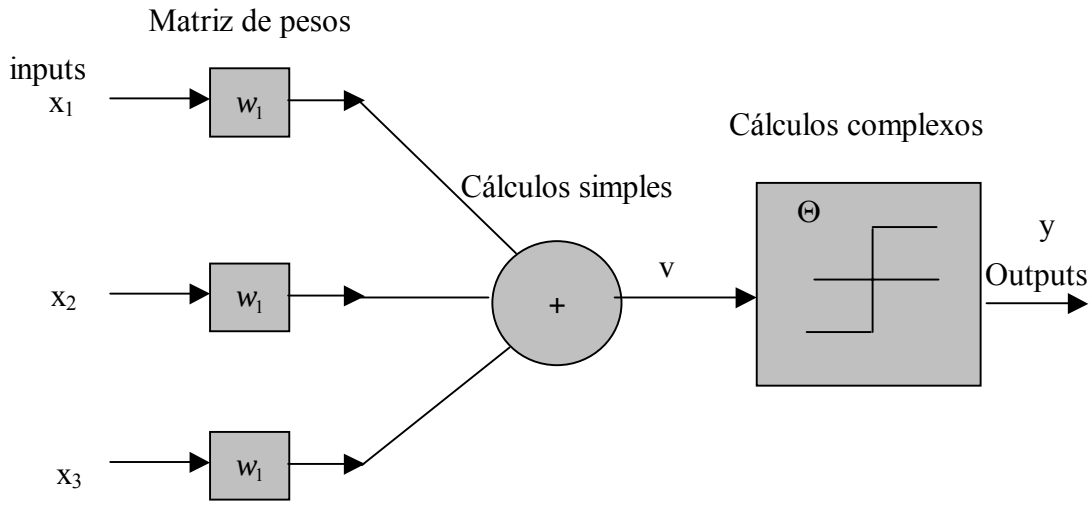
2.18.1 Redes Neurais Artificiais

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 unidas por conexões.

As Redes Neurais geralmente constroem superfícies complexas, utilizando equações algébricas e funções de ajuste, 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.

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. A figura 2.8 ilustra uma rede neural simples.

FIGURA 2.8 – MODELO DE UMA REDE NEURAL SIMPLES -PERCEPTRON



Para construir um modelo neural, primeiramente deve-se adestrar a rede em *dataset* de treinamento e então usar a rede já treinada para fazer predições.

Informações mais detalhadas e o histórico da técnica serão descritos no capítulo 4.

2.18.2 Algoritmos Genéticos (Ags).

Os algoritmos Genéticos realizam a tarefa de classificação de dados baseados na analogia com os processos de seleção natural e genética evolucionária.(Goldenberg, 1989).

A essência da técnica consiste em manter uma população de indivíduos (cromossomos), os quais representam possíveis soluções para um problema. A melhor solução é obtida através de um processo de seleção competitiva ”(Herrera, 1996).

Os Ags foram idealizados por John Holland (1975) tendo sido fundamentados nos

princípios da genética humana populacional.

Segundo esse princípio, a variabilidade entre indivíduos em uma população de organismos que se reproduzem sexualmente, é produzida pela mutação e pela nova combinação genética. Esses algoritmos "compõem uma estratégia de busca e otimização global que tem se mostrado extremamente útil na solução de problemas complexo" (Bäck, Hammel e Schwefel, 1997).

Os Ags pertencem à classe dos algoritmos probabilísticos, mas não possuem característica aleatória, pelo contrário, pois dirigem a busca para regiões do espaço de busca onde é 'provável' que os pontos ótimos estejam.

As pesquisas de Darwin, Lamark e Mendel, entre outros pesquisadores da área biológica, servem de alicerce e norteio para as implementações computacionais baseadas nas teorias evolucionárias. A execução de um Algoritmo Genético se dá da seguinte forma:

a) Uma possível solução é representada por meio de uma seqüência de símbolos, denominada cromossomo, sendo essa representação geralmente feita usando-se a representação binária, ou seja, {0,1}.

b) Em seguida gera-se uma população inicial por processos aleatórios ou heurísticos.

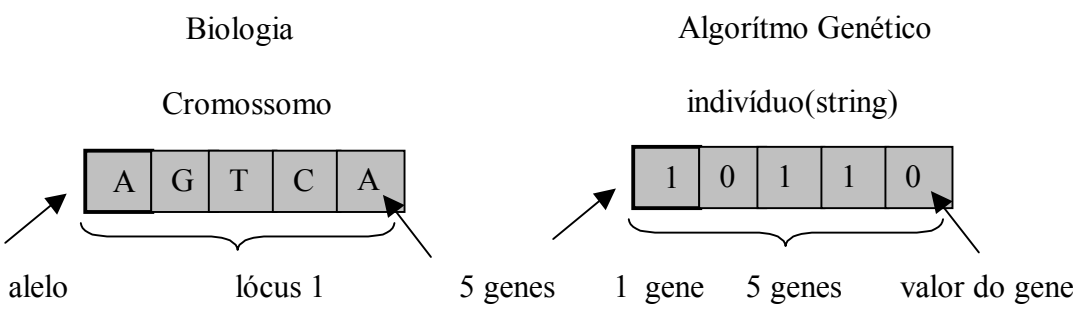
c) Cria-se uma função de avaliação, que na maioria dos casos, é a própria função objetivo do problema, para se avaliar o nível de aptidão de cada cromossomo, é também denominada de função de *fitness.*

d) Na seqüência, na tentativa de melhorar a resposta pode-se trabalhar com operadores genéticos: a mutação, o *crossover* e a clonagem.

e) Finalmente, antes de resolver o problema deve-se definir alguns parâmetros, tais como: critérios de parada, tamanho da população, taxa de *crossover*, taxa de mutação, intervalo de geração, roleta (avalia o *fitness* do cromossomo e o eletismo).

A figura 2.9 faz a analogia entre o cromossomo biológico e o indivíduo (*string*) do Algoritmo Genético.

FIGURA 2.9 – ANALOGIA ENTRE O CROMOSSOMO BIOLÓGICO E O STRING



FONTE: AUTOR

2.18.3 Algoritmo Colônia de Formigas

Um algoritmo de otimização de colônia de formigas é também conhecido como ACO (*Ant Colony Optimization*). O ACO é um algoritmo baseado em agentes que simulam o comportamento natural de formigas, inclusive mecanismos de cooperação e adaptação e para tanto estão baseados nas seguintes noções:

Cada caminho escolhido por uma formiga é associado a uma solução candidata para resolver o problema.

Quando formigas seguem um caminho depositam nesse caminho uma quantidade de feromônio que é proporcional à qualidade da solução candidata.

No momento de optar por dois ou mais caminhos a probabilidade da formiga optar por aquela que possui mais feromônio é maior.

Desta forma a tendência natural é ao longo do tempo as formigas optarem por um caminho mais curto para solução do problema, e que se não é o ótimo está muito próximo dele.

Os ACO geralmente são utilizados para descoberta de regras de classificação e a qualidade de uma regra é representada por Q e calculada pela fórmula:

$$Q = \frac{TP}{TP + FN} * \tag{2.1}$$

onde:

TP (verdadeiros positivos): número de casos cobertos pela regra que tem a classe predita pela regra

FN (falsos negativos): número de casos que não estão cobertos pela regra, mas tem

a classe predita pela regra.

Os valores de Q serão sempre $0 \leq Q \leq 1$ e quanto maior for Q, maior será a qualidade da regra.

Atualiza-se o feromônio para um termo $_{ij}$, com a fórmula:

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \tau_{ij}(t) * Q \quad \forall ij \in R, \tag{2.2}$$

onde:

$\tau_{ij}(t)$ é a quantidade de feromônio no tempo t para um termo(caminho) ij.

$\tau_{ij}(t+1)$ é a nova quantidade de feromônio no termo ij, para o tempo t+1, atualizada pela qualidade da regra Q.

R: é o conjunto de condições que acontecem na regra construída pela formiga na iteração t.

Uma abordagem interessante sobre o uso de um ACO foi desenvolvida por Parpinelli, Lopes e Freitas (2002) e foi denominado *Ant-Miner*.

O *Ant Miner* utiliza-se de regras de classificação do tipo:

SE<termo 1>E<termo 2>E.....>então<classe>

Utilizando-se desta regra o *Ant-Miner* busca descobrir uma lista de regras de classificação. A cada iteração do algoritmo uma regra é descoberta e se casos são corretamente cobertos por essa regra são então retirados do conjunto de treinamento. O processo é então repetido iterativamente enquanto o número de casos descobertos for menor que o limite estabelecido pelo usuário. Resumindo em cada repetição três passos são observados: construção da regra, poda (retirada dos casos cobertos pela regra) e a atualização do feronômio.

2.18.4 Árvores de Decisão

Nesta técnica escolhe-se a variável que se quer avaliar e o *software* procura as mais correlacionadas e monta a árvores com várias ramificações. As Arvores de Decisão (A.D.) são meios de representar resultados de *Data Mining* na forma de árvore, e que lembram um gráfico organizacional horizontal (organograma). Dado um grupo de dados com numerosas linhas e colunas, uma ferramenta de árvore de decisão pede ao usuário do *software* para definir em sua base de dados, qual será o atributo meta (objeto de saída) e então mostra o único e mais importante fator correlacionado com aquele objeto de saída como o primeiro ramo ou nó da árvore de decisão. Os outros atributos preditores subseqüentes são

classificados como nós, do nó anterior, formando gradativamente a árvore.

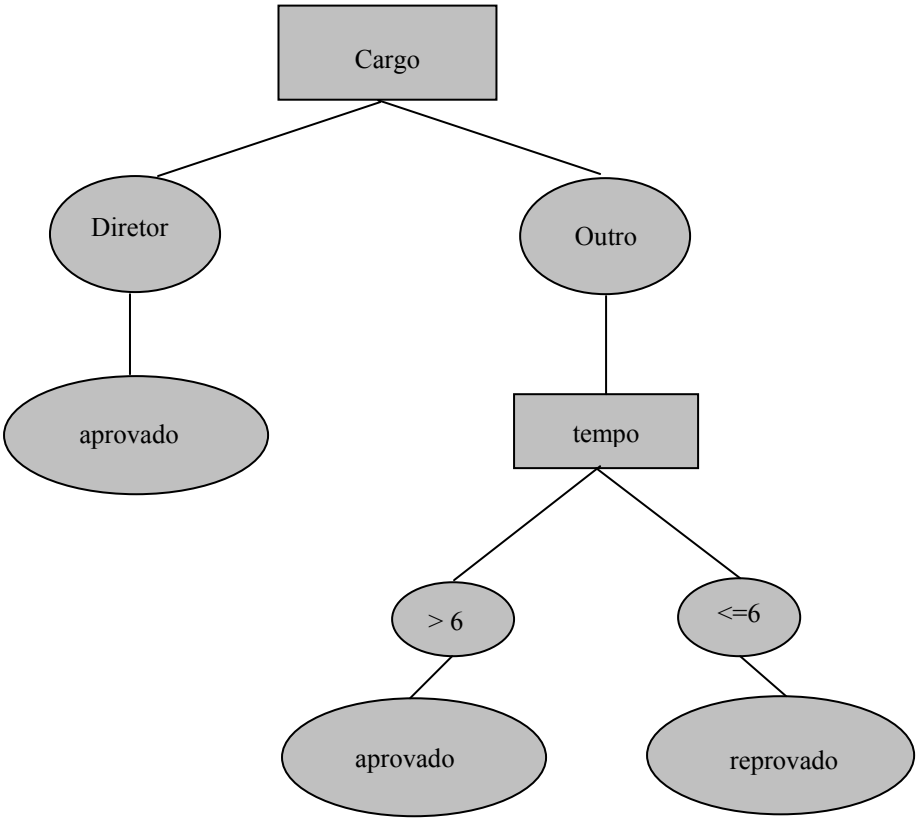
Ferramentas interessantes em árvore de decisão devem permitir ao usuário explorar a árvore conforme sua necessidades e vontades e também propiciar ao usuário que encontre grupos alvos, de maior interesse e aprofundar-se mais nesses grupos.

As árvores de Decisão são utilizadas muitas vezes em parceria com a técnica de Indução de Regras, mas diferenciam-se desta pelo formato com que os resultados são apresentados. Nos resultados da técnica de AD existe a priorização, ou seja, a regra mais importante é apresentada no primeiro nó e as regras menos relevantes ficam para os nós seguintes.

Uma das melhores características da técnica AD é a facilidade de manipulação, aliada a comunicação visual da árvore que facilita a compreensão do usuário.

A figura 2.10 mostra um exemplo de uma pequena árvore de decisão e seus componentes.

FIGURA 2.10 - MODELO DE UMA ÁRVORE DE DECISÃO



FONTE: AUTOR

2.18.5 Regras de Associação

Os problemas que envolvem associação geralmente são solucionados através da técnica de Regras de Associação. Genericamente, uma regra de associação é representada pela notação $X \Rightarrow Y$, ou seja, X implica em Y, onde X e Y são conjuntos distintos.

O objetivo desta técnica é representar, com determinada certeza, uma relação existente entre o antecedente e o conseqüente de uma regra de associação. A associação é uma tarefa descritiva, pois visa identificar padrões em dados históricos.

Um exemplo típico de Regras de Associação sé construído quando utilizamos uma cesta de compra. O objetivo é saber se determinado produto X implica na compra do produto Y. Esta implicação é avaliada através de dois fatores: suporte e confiança.

O suporte de uma regra representa o percentual das transações em que a regra acontece em relação ao total de transações. A confiança não trabalha com todas as transações, apenas com as que possuem o antecedente da regra. Assim a confiança é a razão entre o número de vezes em que o conseqüente da regra aparece, pela quantidade dessas transações.

Um bom algoritmo de extração de regras deve gerar regras que possuam suporte e confiança especificados pelo usuário e as regras podem ser compostas de um ou mais itens.

Se a base de dados for muito grande, muitas regras podem ser criadas. Cabe ao usuário saber selecionar as melhores e que serão importantes na tomada de decisão.

2.18.6 Análise de Agrupamento.

A tarefa de agrupamento é descritiva, ou seja, ela visa identificar padrões em uma massa de dados. A principal diferença entre a classificação e o agrupamento é que nessa última as classes não são previamente definidas. A idéia é que o algoritmo de agrupamento identifique automaticamente comportamentos similares em uma base de dados, dividindo a massa de informações em grupo.

Após o processo de agrupamento o analista deve estudar os padrões identificados procurando verificar se esses podem ser transformados em conhecimento estratégico. O agrupamento não responde porque os padrões existem, apenas os identifica.

As técnicas de agrupamento mais conhecidas são: o agrupamento por Particionamento e agrupamento Hierárquico.

No Particionamento o algoritmo das *K-means* é o mais utilizado e divide o grupo total de itens em subgrupos. O *dataset* é tratado como um vetor e, cada informação é

considerada uma componente vetorial. Se o banco de dados não for numérico a técnica não funciona, portanto os dados qualitativos devem ser transformados em variáveis numéricas, pois o algoritmo trabalha com distâncias entre os pontos. Pontos próximos formam um grupo. Para plotagem dos dados, lança-se mão de uma função de distância e as mais utilizadas são a função Euclidiana e a função *Manhattan*. Faz-se necessário também definir previamente o número de grupos a ser criado e esse número é denomidado de K, daí o nome *K*-médias. Inicia-se então com o algoritmo dividindo o *dataset* em K grupos e plotando um ponto chamado centróide (*mean*) no meio de cada grupo. Na seqüência reposiciona-se os centróides de acordo com sua distância em relação aos outros pontos do grupo. Com o centróide reposicionado, os grupos são novamente plotados. Em seguida os centróides são novamente recalculados e o processo se repete até que os grupos estejam bem definidos.

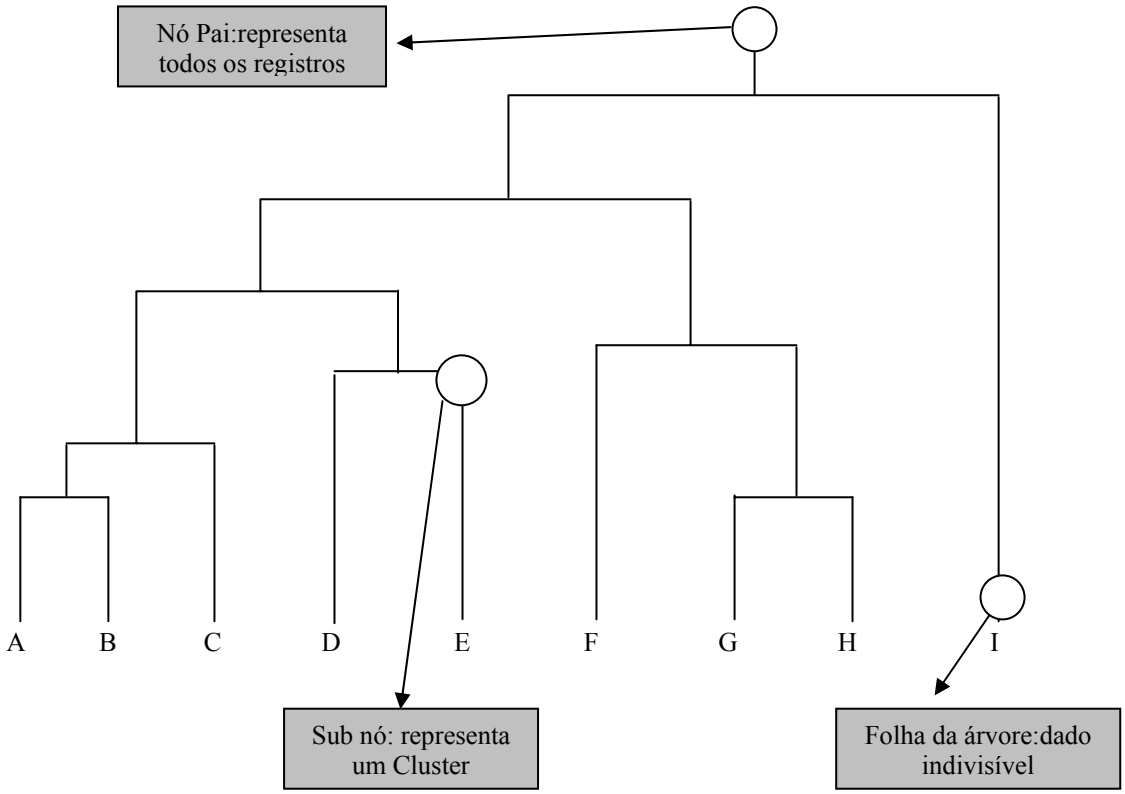
No Agrupamento Hierárquico o algoritmo mais utilizado é o HAC. Essa técnica trabalha de duas formas:

a) processo divisivo: começa com um cluster único e vai particionando-o em clusters menores, em processo iterativo.

b) Processo aglomerativo: faz o processo ao contrário, inicia em cluster pequeno e vai se agrupando em clusters maiores.

O resultado é uma árvore de grupos chamada de dendrograma. A figura 2.11 mostra um exemplo de dendrograma.

FIGURA 2.11 – DENDROGRAMA FORMADO POR 9 FOLHAS E 1 NÓ PAI.



FONTE: REVISTA SQL MAGAZINE NR 10 – ANO 1

2.18.7 Regressão

Regressão também conhecida como estimativa é considerada uma tarefa preditiva, pois seu objetivo é prever um valor numérico desconhecido a partir de alguns atributos conhecidos, utilizando uma massa de dados histórica como modelo.

As técnicas mais comuns de estimativa são baseadas nos mesmos métodos da classificação, ou seja, utilizam árvores de decisão. Em outras palavras, a idéia é a geração de um modelo que possam estimar o valor numérico de determinado atributo.

2.18.8 Regressão Logística.

A técnica de Regressão Logística, que trata do ajuste do Modelo Logístico, é em geral, utilizada para tratar problemas relacionados a dados dicotômicos em várias áreas do conhecimento. Em *Data Mining* é interessante saber qual a probabilidade de um indivíduo pertencer a um determinado grupo.

Este modelo estabelece uma relação entre a probabilidade de ocorrência dos resultados de uma variável resposta dicotômica, que geralmente é representada pelos termos sucesso e fracasso, e variáveis explicativas categóricas ou contínuas.

A idéia básica consiste em estabelecer uma relação linear entre variáveis explicativas e uma transformação denominada logito (*logit*), da variável resposta.

Este modelo é representado por (Hosmer e Lemeshow,1989; Arminger et al. 1997)

$$\log\left[\frac{\mathrm{P}\{\mathrm{Y}(\mathrm{x})=1\}}{\mathrm{P}\{\mathrm{Y}(\mathrm{x})=0\}}\right] = \beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_p x_p, \tag{2.3}$$

onde:

$P\{Y(x)=1\}$ representa a probabilidade de sucesso para a variável resposta e $P\{Y(x)=0\}$ representa a probabilidade de fracasso, β_0 denota o intercepto da regressão e $x' = (x_1, x_2, \ldots\ldots, x_p)$ é um vetor de variáveis explicativas com coeficientes $\beta_1, \beta_2, \ldots, \beta_p$).

Desta forma, a probabilidade de sucesso para a variável resposta é dada por:

$$p(x) = P\{Y(x)=1\} = \frac{e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}} \tag{2.4}$$

O ajuste do modelo pode ser feito pelo método da verossimilhança ou dos mínimos quadrados. Está técnica será discutida com maior destaque no capítulo IV.

2.18.9 Classificação Bayesiana

A Classificação Bayesiana é uma técnica não supervisionada de classificação de dados. Não divide os dados em classes, porém define uma função probabilística pertinente a cada classe de dados e é fundamentada no teorema de Bayes para probabilidades condicionais:

$$P\{A/B\} = \frac{P\{A \cap B\}}{P\{B\}} \tag{2.5}$$

Na abordagem Bayesiana os principais conceitos envolvidos são a probabilidade a *priori* e a probabilidade a *posteriori*. A probabilidade de um evento ocorrer pode ser refinada de posse de outros dados históricos. Desta forma a probabilidade a *priori* seria modificada através de um fator que varia com a análise dos dados históricos. Esta nova probabilidade é denominada probabilidade a *posteriori* ou probabilidade condicional de um evento ocorrer dado à época ou o local, por exemplo.

Este fator multiplicativo pode ser determinado a partir de dados históricos que adquire o formato de uma equação como segue:

$$P\{A/B\} = P\{A\} * fator(B) \tag{2.6}$$

onde

$P\{A\}$ é a probabilidade a *priori* e

$P\{A/B\}$ é a probabilidade a *posteriori*.

Algumas evidências podem aumentar a probabilidade se o fator for maior do que 1, enquanto que outras podem diminuir a probabilidade se o fator for menor do que 1.

Essa técnica de classificação pode ser usada para predizer a probabilidade de um novo indivíduo, elemento ou população ser encontrado em uma determinada classe.

2.18.10 Análise Discriminante

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 obter 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. Para usar a técnica utiliza-se a Função Discriminante de Fisher para dois grupos ou vários grupos. A técnica aplicada para dois grupos é descrita na seqüência.

Considera-se inicialmente, duas populações π_1 e π_2. A função discriminante de Fisher é constituída sem assumir qualquer forma paramétrica para os grupos, ou seja, sem assumir a existência de uma função de probabilidade associada a cada grupo. A idéia de Fisher é procurar por uma regra, sensível o suficiente, que possa discriminar entre as duas populações, de tal modo que as observações multivariadas $\underline{X}$, possam ser transformadas em observações univariadas Y, tal que os Y's nas populações π_1 e π_2 estejam o mais distante possíveis. Trabalha-se então com a função linear $\underline{C}'\underline{X}$ a qual deve maximizar a razão entre o quadrado das distâncias entre as médias das populações univariadas e a variância de Y, ou

seja $\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2}$.

Desta forma a função linear $\underline{C}'\underline{X}$ é denominada Função Linear Discriminante de Fisher. Na função tem-se que o vetor $\underline{C}$ é o auto-vetor da matriz $W^{-1}B$ que corresponde ao máximo autovalor, onde W e B são as matrizes das somas de quadrados e produtos cruzados dente grupos e entre grupos respectivamente. As médias amostrais $\bar{x}_i$ terão escores $C'\bar{x}_i$ e , no caso de apenas dois grupos , $\underline{C} = W^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2})$. Após a função discriminante ter sido determinada, um novo objeto pode ser alocado a uma das duas populações levando-se em conta o escore discriminante $\underline{C}'\underline{x} = (\underline{\bar{x}_1} - \underline{\bar{x}_2})'W^{-1}\underline{x}$. O objeto com variáveis medidas x_0 é alocado na população cujo escore médio é próximo a $\underline{C}'\underline{x}_0$. Sendo assim aloca-se o objeto ao grupo π_1 se $(\underline{\bar{x}_1} - \underline{\bar{x}_2})'W^{-1}\{\underline{x}_0 - \frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_2})\} > 0$ e ao grupo π_2 caso contrário. Considerando-se que as duas populações possuem matriz de variâncias-covariâncias comum, pode-se substituir a matriz W pela matriz S que é a combinação de dois estimadores da matriz variância-covariância e desta forma a alocação do novo objeto obedece o seguinte critério: aloca-se o objeto ao grupo π_1 se $(\underline{\bar{x}_1} - \underline{\bar{x}_2})'S^{-1}\{\underline{x}_0 - \frac{1}{2}(\underline{\bar{x}_1} + \underline{\bar{x}_2})\} > 0$, caso contrário aloca-se no grupo π_2.

2.18.11 Técnicas de Visualização

São técnicas utilizadas para sumarizar grandes quantidades de dados, em muitas situações, suficientes para extração de conhecimento, descoberta de padrões, tendências e relações dentro de um grupo de dados.

Por outro lado deve-se salientar que nenhuma técnica analítica é executada, pois o usuário dos *softwares* de visualização restringe-se simplesmente em visualizar gráficos, mapas ou tabelas estatísticas básicas.

O método de visualização escolhido depende do tipo de conjunto de dados a ser analisado. Segue alguns métodos mais comuns utilizados na técnica de visualização: diagramas de associação, tabelas, matriz de associação, mapas, gráficos baseados em proporção em ícones, diagramas hierárquicos, diagramas híbridos, entre outros.

O quadro 2.1 na página seguinte, apresenta o resumo das tarefas em *Data Mining*, a função da tarefa, a técnica utilizada e a ferramenta (*software*) disponível para executá-la.

QUADRO 2.1 - CLASSIFICAÇÃO DAS TÉCNICAS EM *DATA MINING*

Tarefa	**Função da Tarefa**	**Técnica**	**Algoritmos**
Associação	Descrição	Regras de Associação	Apriori, DHP,ABS, Sampling,PARTITION,AQ. Lógica Fuzzy,WEKA
Classificação	Predição	Colônia de Formigas	Ant-Miner
		Árvores de Decisão	ID3,ID4,J48,CART,C4.5
		Algoritmo Genético	DeCan, RAGA.
		Regressão Logística[4]	Mínimos Quadrados, Máxima Verossimilhança.
		Classificação Bayesiana.	Teorema de Bayes (Probabilidade Condicional)
	Sumarização	Visualização	IBM, SAS,AVS,Visualization Edition.
Estimativa	Predição	Redes Neurais [5]	(MLP, Percepton, Adeline e Madeline, Backpropagation, TANAGRA).
		Regressão Linear	Regressão estatística(M5, CART).
		Análise de Discriminante[6]	Função Discriminante de Fischer
Clusterização	Descrição	Análise de Agrupamento	K-means, HAC

FONTE: AUTOR.

2.19 APLICAÇÕES DE *DATA MINING*.

[4] A técnica será utilizada no desenvolvimento do trabalho.

[5] Idem.

[6] Idem.

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, controle de qualidade, marketing e medicina (LOUZADA e DINIZ, 2002).

2.19.1 *Data Mining* em Comércio.

Utilizando o imenso banco de dados disponíveis em setores comerciais pode-se utilizar o *Data Mining* para alocação de novas filiais, ou seja, determinar a localização para abertura de uma nova filial otimizando vendas.

Grandes grupos supermercadistas utilizam *Data Mining* para estudar o comportamento de compra de seus clientes. Através do cadastramento de clientes com um cartão específico, que utilizado no momento da compra, identifica as características pessoais do cliente, tais como, sexo, idade, estado civil, etc, e as características dos produtos adquiridos. A análise dos dados pode motivar novos clientes ou ainda manter a clientela padrão, com promoções, eventos, vendas casadas, etc.

2.19.2 *Data Mining* em Finanças.

Bancos, instituições financeiras e entidades de proteção ao crédito, vêm utilizando técnicas de *Data Mining* em seus bancos de dados para criar sistemas de avaliação de crédito, objetivando predizer se o cliente será adimplente ou inadimplente.

2.19.3 *Data Mining* em Seguros

Grandes companhias de seguro apresentam perdas devido ao cancelamento de apólices e custos gerados para obtenção de novos clientes. Ferramentas de *Data Mining* podem ser utilizadas analisando as características dos clientes, predizendo quem cancelaria as suas apólices com uma certa margem de segurança. Os resultados obtidos oferecem informações riquíssimas aos administradores que podem tratar os clientes propícios ao cancelamento de forma diferenciada, reduzindo as apólices canceladas e os custos das empresas.

2.19.4 *Data Mining* em Medicina

As ferramentas de *Data Mining* estão sendo muito utilizadas na medicina. Através da análise do banco de dados (histórico) dos pacientes consegue-se identificar correlações entre variáveis imperceptíveis a olho nu, auxiliando no sucesso de órgãos transplantados, redução de efeitos de quimioterapia e análise da corrosividade da pele, provocada pelo uso de novos produtos que serão lançados no mercado.

2.19.5 *Data Mining* no Governo

O Governo dos Estados Unidos utiliza-se de *Data Mining* para identificar padrões de transferência de fundos internacionais que se parecem com lavagem de dinheiro, varre banco de dados buscando compradores de explosivo, armas e munições, na tentativa de diminuir o índice de crimes e atentados terroristas.

2.19.6 *Data Mining* em Marketing

A avaliação do hábito de usuários de computador, através de ferramentas de *Data Mining* pode gerar o Marketing Eletrônico, criando e-mails específicos para determinados grupos de consumidores e oferecendo *sites* direcionados as necessidades de consumo desses grupos.

2.19.7 *Data Mining* no Vestibular.

Utilizando técnicas de *Data Mining*, um programa de obtenção de conhecimentos, após analisar milhares de alunos matriculados na PUC-RJ, constatou que se o candidato trabalha, é do sexo feminino e teve aprovação com boas notas, não efetivará a matrícula pois é muito provável tenha sido aprovado também em uma universidade pública.

2.19.8 *Data Mining* em Telecomunicações

Técnicas em *Data Mining* podem ser utilizadas para evidenciar os hábitos dos usuários de telefones celulares, contribuindo para diminuir a clonagem, crimes contra a telefonia. Quando um telefonema é realizado fora dos padrões o *software* faz uma chamada para o cliente confirmando se foi ou não uma fraude.

2.20 TRABALHOS RELACIONADOS NA ÁREA DE DATA MINING

CHIARA (2003) utilizou *Data Mining* na identificação de padrões entre usuários da internet, que na seqüência de cliques em seus computadores vão deixando informações preciosas sobre suas personalidades e preferências. Esses dados podem ser armazenados em um *Data Warehouse* para serem analisados por técnicas de *Data Mining*. Neste trabalho são discutidas e analisadas algumas das técnicas utilizadas para atingir esse objetivo. É proposta uma ferramenta onde os dados dessas seqüências de cliques são mapeados para o formato atributo-valor utilizado pelo Sistema *Discover*, (um sistema sendo desenvolvido no Laboratório do Instituto de Ciências Matemáticas e de Computação da USP-SP) para o planejamento e execução de experimentos relacionados aos algoritmos de aprendizado utilizados durante a fase de Mineração de Dados do processo de descoberta de conhecimento em bases de dados.

CANUTO e GOTTGTROY (1997), apresentaram proposta Sistema Especialista Híbrido (SAGRI) utilizado na agricultura, que tem como principal objetivo o apoio e aconselhamento aos técnicos pesquisadores e, em especial o agricultor nas etapas do processo produtivo, melhorando a utilização dos recursos naturais disponíveis. O sistema SAGRI utiliza grandes bases de dados que dificultam, devido ao volume, a atualização e eficiência na tomada de decisão. Nesse trabalho foram utilizadas técnicas de *Data Mining*, em vários pontos do sistema SAGRI e detectado a técnica mais adequada para a realização das tarefas específicas. Desta forma foi utilizado Redes Neurais na fase de pré–processamento dos dados de forma a organizá-los e depurá-los. O objetivo da utilização do *Data Mining* foi o de relacionar características do solo e culturas. Com a descoberta do conhecimento, é facilitado o processo de definição de qual solo é mais adequado para uma cultura que até o momento não tinha sido introduzida em uma determinada região.

AZEVEDO (2002), escreveu artigo a respeito das oportunidades de se construir melhor relacionamento com clientes, aumentar freqüência de compras e conseqüentemente aumentando lucros. Em operações de venda é essencial manter as transações realizadas o que inclui histórico de compras com produtos, preços e datas, os contatos com a empresa e as respostas dadas aos estímulos de *marketing*. Portanto a quantidade de dados sobre clientes é enorme e suas análises ficam facilitadas com a utilização de ferramentas de *Data Mining*. Reunindo estatística e inteligência artificial**,** as ferramentas buscam automaticamente padrões

nos dados, gerando modelos de comportamento. Cada cliente é classificado com uma nota (escore) que representa a probabilidade de resposta positiva ao estímulo posto no modelo.O *Data Mining* auxilia o profissional de marketing a focar mais precisamente suas campanhas, alinhando-as às necessidades, desejos e atitudes de clientes e prospectos.

SOARES e NADAL (1999), apresentaram trabalho utilizando técnicas de *Data Mining* na tentativa de definir critérios para análise de sinais de ECGA (Eletrocardiograma Ambulatorial) para diagnóstico de integridades cardíacas, onde alterações no segmento ST dos eletrocardiogramas refletem a ocorrência de crises istêmicas. Foram utilizadas no trabalho a redução de dimensionalidade através da análise de componentes principais para extração de parâmetros e posteriormente foi empregado Redes Neurais Artificiais (RNA), do tipo *feedforward*[7], treinadas com o algoritmo Leveenberg- Marquardt[8] para a tarefa de classificação de padrões. Os resultados foram considerados satisfatórios.

SILVA et all. (2003), escreveu artigo que descreve um estudo sobre o uso de *Data Mining* em um ambiente *web*[9]. Técnicas de processo de descoberta de conhecimento foram aplicadas com o intuito de investigar a relevância das informações obtidas por meio da análise dos padrões de navegação de usuários em *web sites*[10] de uma empresa provedora de acesso à Internet, descritos em arquivos de log[11] de um servidor *Web*. A partir disto, medidas foram sugeridas para um melhor aproveitamento e eficácia do processo.Para efetuar a análise dos dados, foi escolhida a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*), Para a realização deste estudo, foi utilizado o algoritmo *Apriori*[12], implementado pela

---

[7] Redes feedforward:

[8] Leveenberg- Marquardt: é um algoritmo de aproximação que utiliza o método de Gauss-Newton modificado pela introdução de um parâmetro μ que funciona como fator de estabilização no treinamento da rede.

[9] Web: termo associado a Internet para definir a rede mundial de computadores.

[10] Web site: local específico na rede, endereço de internet.

[11] Log: anotações das atividades ocorridas no computador ou entre dois computadores.

[12] Detalhes sobre o algoritmo *Apriori* podem ser encontrados em (AGRAWAL et al., 1994).

ferramenta e que faz uso de regras de associação.

Considerando-se a mineração de dados realizada sobre os dados de acessos ao *web site,* registrados no período analisado e tendo como atributos de análise o período, tipo de internauta e páginas acessadas, a aplicação do algoritmo *Apriori*, por meio da ferramenta Weka[13], permitiu extrair várias regras de associação, o que propiciou a identificação de padrões de comportamento de internautas ao navegarem pelo *web site* da empresa.

Ao ter conhecimento da freqüência com que determinadas seções do web *site* são acessadas e quais são os serviços mais procurados, a gerência da empresa pôde descobrir o perfil de seus usuários e, com base nisso, ofertar serviços e atendimento personalizado.

GOMES et all. (2003), demonstrou em seu trabalho a viabilidade de uso de *softwares* livres de *Data Mining* (*Mining Tools e Weka*) para classificação de dados. Os dados utilizados nesse trabalho foram obtidas utilizando indicadores do censo demográfico de 2000 e do sendo escolar de 2000, publicada no sítio do INEP, que por sua vez apresenta dados de IDH- Índice de Desenvolvimento Humano relacionado com dados gerados pelo censo educacional. Os resultados alcançados indicaram que o IDH não apresentou uma relação forte com os indicadores de desempenho e investimentos em educação.

RODRIGUES (2000), elaborou trabalho de mestrado onde traçou um panorama das diversas técnicas de dada mining sob o ponto de vista do usuário. Para tanto utilizou uma matriz de classificação tridimensional com três informações principais de cada técnica estudada: tipo de problema, método de abordagem e área de aplicação.

O objetivo da matriz é prover informação através dos cruzamentos entre as três dimensões de classificação. Fixando um item de uma dimensão, pode-se percorrer a matriz em duas dimensões, analisando onde e como as dimensões interagem.Para análise da matriz de classificação fixa-se dimensão de área de aplicação. Para cada área de aplicação tem-se então uma matriz bidimensional com tipos de problemas e métodos de abordagem.Desta forma, tem-se ao final de cada análise uma varredura para cada área de aplicação considerando a aplicação de tipos de problemas e quais metodologias são utilizadas para solucioná-los.

Uma das maiores contribuições da matriz de classificação está em sua utilização

para auxiliar no processo de tomada de decisão de um projeto de Descoberta de Conhecimento, pois permite que a técnica de *Data Mining* selecionada seja a mais adaptada possível ao problema, poupando tempo e custos.

QUEIROZ et all. (2002), escreveram artigo sobre a Interface entre Homem e o Computador (IHC), com o objetivo de avaliar, através de extração de regras,o uso das interfaces e o aprendizado promovido. Para desenvolvimento do trabalho foi utilizado um *software*, o Cabri Geometrè[14]. Cada usuário ao utilizar o *software* absorvia ou não o conhecimento com características pré-definidas. Essas características individuais formaram o banco de dados que depois de trabalhado foi analisado pela ferramenta Weka, sendo utilizado para tal, vários algoritmos com destaque para o J48-PART[15].

PAULA et all.(2000), utilizaram em seu trabalho, redes neurais artificiais de múltiplas camadas para auxiliar na classificação em diferentes graus, de possível sonegação fiscal, os contribuintes de um ramo específico de atividade econômica em região da Zona da mata de Minas Gerais. Os dados foram gerados com indicadores setoriais e individuais, baseados em informações fiscais e contábeis. Os resultados obtidos conseguiram classificar as empresas estudadas de acordo com o grau de sonegação fiscal com exatidão entre 97,8% e 100% .

---

[13] Weka: *software* livre encontrado em www.cs.waikato.ac.nz/~ml/weka/index.htm.

[14] Cabri Geometrè: *software* livre para ensino de geometria encontrado em http://www.cabri.com/v2/pages/us/index.php

[15] O Algoritmo J48 é uma implementação da ferramenta Weka para a técnica C4.5 *decision tree ou(um tipo de árvore de classificação)*

CAPÍTULO III

3 TÉCNICAS UTILIZADAS.

Este capítulo contém as técnicas de *Data Mining* efetivamente utilizadas neste trabalho, com um detalhamento maior de cada técnica, os modelos matemáticos e métodos utilizados e os erros existentes na aplicação das técnicas na resolução dos problemas.

As técnicas de *Data Mining* utilizadas foram escolhidas conforme a necessidade de atingir os objetivos previstos no capítulo I, deste trabalho. Para tanto se faz necessário lembrar o objetivo principal do trabalho: "o objetivo principal do presente trabalho 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 pré-dizer se um novo indivíduo pertence ao grupo dos satisfeitos ou insatisfeitos, e desta forma, diminuir a rotatividade de funcionários".

Portanto, a predição é a tarefa a ser realizada. Após análise de diversas técnicas de predição, três delas foram escolhidas, por se adequarem melhor ao banco de dados disponíveis 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 realizado uma Análise Fatorial, técnica multivariada, utilizada para reduzir a dimensionalidade do problema. Desta forma as técnicas usadas foram as seguintes:

- Técnica de Redução: Análise Fatorial
- Técnicas de predição:
    a) Redes Neurais Artificiais
    b) Análise de Discriminante de Fischer
    c) Regressão Logística.

No capítulo II as técnicas foram descritas de forma resumida, na seqüência seguem detalhamentos de cada técnica e também detalhes sobre a técnica Análise Fatorial, realizada previamente objetivando reduzir a dimensionalidade do problema.

3.1 REDUÇÃO DE DIMENSIONALIDADE

Um problema bastante comum que surge quando se utiliza técnicas de *Data Mining* é a necessidade de reduzir a dimensionalidade do problema estudado. O processo de redução é utilizado quando o número de variáveis estudadas é muito grande. Para reduzir a dimensão do vetor original pode-se criar um novo vetor cujas componentes são combinações lineares das variáveis originais.

A redução da dimensionalidade do vetor de variáveis originais pode ser feita segundo dois critérios (Draper e Smith, 1998):

- Conhecimento informal do especialista utilizando a própria vivência em torno do problema, ou seja, conhecimento empírico.
- Conhecimento Científico através de critérios estatísticos, conhecidos como técnicas de seleção de variáveis.

A criação de um novo vetor de variáveis de menor dimensão, cujos componentes são combinações lineares das variáveis originais, pode ser conduzida via uma técnica de análise multivariada, conhecida como Análise de Componentes Principais (Anderson, 1984 e Johnson, 1982).

Outro critério interessante e muito utilizado para reduzir a dimensionalidade dos problemas é a aplicação da Análise Fatorial, que além de poder utilizar o critério das componentes principais para estimar os pesos fatoriais, possibilita a visualização de grupos que agregam variáveis e são chamados de fatores, oferece o cálculo das comunalidades[16] e os valores residuais.

Na seqüência segue a descrição do Método das Componentes principais e a Análise Fatorial.

3.1.1 Análise de Componentes Principais

Como foi citado no parágrafo anterior as componentes principais são determinadas através de combinações lineares padronizadas de $\underline{X}$, onde $\underline{X}$ é o vetor das variáveis originais. Uma combinação do tipo $\underline{\ell}'\underline{X}$, onde $\underline{\ell}' = (\ell_1, \ell_2, ...\ell_p)'$ é um vetor (transposto) de constantes, pode ser denotado como uma combinação linear padronizada (CLP) se $\sum_i \ell_i^2 = 1$.

O objetivo principal da técnica Análise de Componentes Principais é encontrar uma CLP das variáveis originais de tal sorte que a variância seja maximizada. A análise de componentes principais busca algumas combinações que possam sumarizar os dados e ao mesmo tempo minimizar a perda de informações. Se um vetor $\underline{x}$ original possui p componentes, e estas componentes são necessárias para produzir a variabilidade total do sistema, grande parte desta variabilidade poderá ser explicada por um número menor k de combinações denominadas de componentes principais. Estas k componentes principais podem substituir as p variáveis originais e o conjunto de dados passa a ter então, n medidas em k componentes principais, k<<p.

Para desenvolver a tarefa de redução de dimensionalidade pode-se partir de duas matrizes de dados originais:

- Matriz de covariância: utilizada quando os dados possuem as mesmas unidades de medida.
- Matriz de Correlação: utilizada quando os dados possuem escalas diferentes de magnitudes como, por exemplo: x_1 de ordem de unidades, x_2 de ordem de milhões, etc.

A matriz preparada utilizada no trabalho enquadra-se no segundo caso, ou seja, o conjunto de variáveis possui magnitudes diferentes, desta forma optou-se por trabalhar com a Matriz de Correlação no momento de reduzir a dimensionalidade do problema. A seguir a técnica é descrita com mais detalhes.

3.1.1.1 Componentes Principais Utilizando a Matriz de Correlação.

De posse da o vetor $\underline{X}$ e suas p componentes, calcula-se primeiramente o vetor de médias ou esperança $\underline{\mu}$, o vetor padronizado $\underline{z}$, a matriz de covariância Σ e utilizando-se da anterior e da matriz desvio padrão $V^{\frac{1}{2}}$,calcula-se a matriz de correlação denominada de ρ.

Desta forma têm-se os seguintes vetores, matrizes e equações[17].

---

[16] Porção da variância de uma variável X, distribuída entre fatores comuns.

[17] As dimensões dos vetores e matrizes já estão adaptadas ao número de variáveis e indivíduos que serão trabalhados no capítulo IV, ou seja, 21 variáveis e 826 indivíduos.

Matriz de dados x de ordem 826 x 21:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & . & . & . & x_{1,21} \\ x_{2,1} & x_{2,2} & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{826,1} & x_{826,2} & . & . & . & x_{826,21} \end{bmatrix}.$$

Vetor de médias $\underline{\mu}$ de dimensão 21:

$$\mu = \mathrm{E(x)} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_{21} \end{bmatrix}$$

Matriz de Covariância:

$$V(\underline{X}) = \Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & . & . & . & \sigma_{1,21} \\ \sigma_{2,1} & \sigma_{2,2} & . & . & . & \sigma_{2,21} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{21,1} & \sigma_{21,2} & . & . & . & \sigma_{21,21} \end{bmatrix}$$

Matriz Desvio Padrão:

$$V^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & . & . & . & 0 \\ . & \sigma_2 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \sigma_{21} \end{bmatrix}$$

Matriz Correlação

$$\rho = \begin{bmatrix} 1 & \rho_{1,2} & . & . & . & \rho_{1,21} \\ \rho_{2,1} & 1 & . & . & . & \rho_{2,21} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho_{21,1} & . & . & . & . & 1 \end{bmatrix}$$

Escore padronizado:

$$z = \frac{(x - \mu)}{(V)^{\frac{1}{2}}}$$

onde:

X = Matriz de dados originais;

μ = E($\underline{X}$) vetor médio estimado por $\overline{\underline{X}}$;

Σ = Matriz Covariância de $\underline{X}$ estimada por S;

$V^{\frac{1}{2}}$ = Matriz Desvio Padrão de $\underline{X}$ estimada por $s^{\frac{1}{2}}$;

z = Equação para padronização e

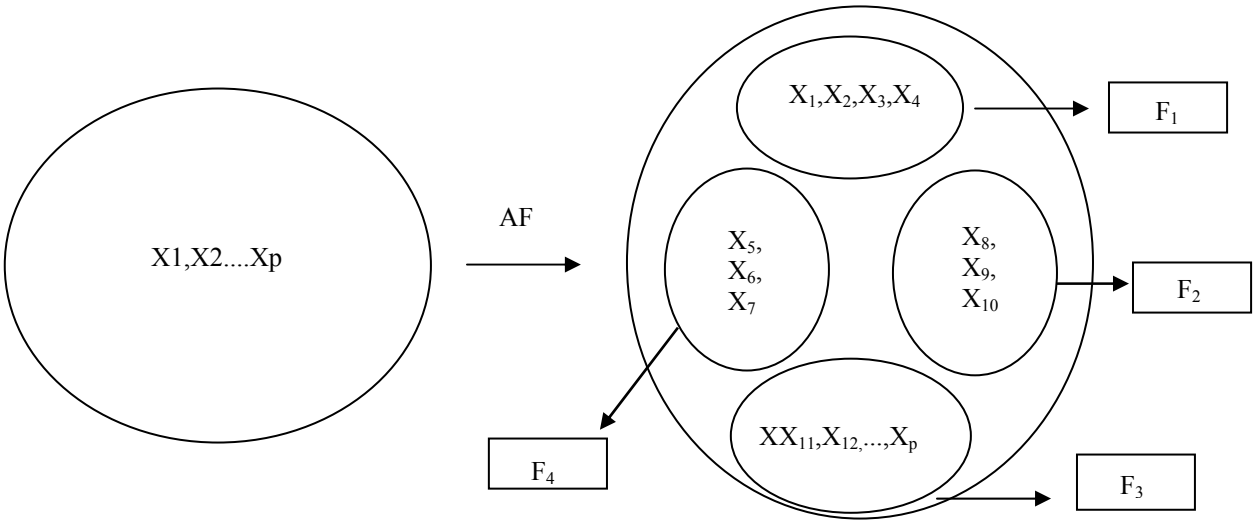
ρ = Matriz Correlação de $\underline{X}$ estimada por R

As componentes principais podem agora ser obtidas através dos autovetores da matriz de correlação ρ. Representa-se por $(\lambda_1^*, e_1^*), (\lambda_2^*, e_2^*), ..., (\lambda_{21}^*, e_{21}^*)$, os pares de autovalores e autovetores da matiz de correlação ρ.Então, tem-se que o i-ésimo componente principal do vetor de variáveis padronizadas $\underline{z}$ é dado por: $Y_i = \underline{e}_i^{*'} \underline{z}, i = 1,2,.....21$. Assim a variância total da população será igual a p e conseqüentemente a proporção da variância total explicada pelo k-ésimo componente principal é dada por: $\frac{\lambda_k}{21}, k = 1,2...21$.

3.1.2 Análise Fatorial

A Análise Fatorial (AF) é uma técnica de análise multivariada que objetiva explicar as correlações existentes entre um conjunto grande de variáveis em termos de um conjunto de poucas variáveis aleatórias não observáveis, denominadas fatores. A AF pode ser mais bem visualizada no esquema da figura 3.1 que se segue.

FIGURA 3.1 ESQUEMA DE AGRUPAMENTO DE VARIÁVEIS EM FATORES (AF)



Quanto mais fortes forem as correlações entre algumas variáveis dentre o grupo inicial, mais nítida é a visualização do fator gerado. Variáveis agrupadas num mesmo fator possuem portanto alta correlação, enquanto que variáveis de fatores distintos possuem baixa correlação.

3.1.2.1 Modelo fatorial ortogonal

Considerando-se $\underline{X}$ o vetor das variáveis originais, com dimensão p, com vetor de média $\underline{\mu}$ e matriz de Covariância Σ, não necessariamente com distribuição normal, tem-se que o modelo fatorial de $\underline{X}$ é linearmente dependente sobre algumas variáveis aleatórias, não observadas $F_1,F_{2,}...,F_{m.}$, (sendo $m < p$, necessariamente caso contrário não haveria ganho com a AF), que são denominados fatores comuns, p fontes de variações aditivas, $\varepsilon_1, \varepsilon_2 ... \varepsilon_p$ são os erros , o modelo de Análise Fatorial pode ser representado com notação matricial como:

$$\underline{X} - \underline{\mu} = L\underline{F} + \underline{\varepsilon} \qquad (3.0)$$

onde:

$\underline{X}$ é o vetor das variáveis originais para cada indivíduo;

$\underline{\mu}$ é o vetor das médias das i-ésimas variáveis;

L é a matriz dos pesos ou carregamento nas i-ésimas variáveis e nos j-ésimos fatores F_j;

$\underline{F}$ é o vetor dos fatores comuns;

$\underline{\varepsilon}$ é o vetor dos erros ou fatores específicos

3.1.2.3 Pesos fatoriais

A interpretação dos fatores de uma AF é feita por meio dos pesos ou cargas fatoriais, que são parâmetros de um modelo de AF que expressam as covariâncias entre cada fator e as variáveis originais. No caso de se utilizar variáveis padronizadas (matriz de correlação), esses valores correspondem às correlações entre os fatores e as variáveis originais. Os pesos ou carregamentos fatoriais são estimados pelo método das componentes principais e serão discutidos com maiores detalhes no capítulo 4.

3.1.2.3 Escores fatoriais

Os escores fatoriais consistem no produto matricial entre os valores observados para as variáveis e os pesos fatoriais. Em muitas aplicações necessita-se estimar o valor de cada um dos fatores para uma nova observação individual $\underline{x} = \lfloor x_1 \; x_2 \; x_3 \ldots x_p \rfloor$. Estes valores dos fatores são denominados de escores fatoriais e serão também discutidos no capítulo 4, durante a Análise Fatorial desenvolvida no presente trabalho, no item 4.8, quando um novo indivíduo será classificado.

3.1.2.4 Comunalidades.

A variância de cada variável X_i é a soma das Comunalidades (h_i^2) com sua respectiva especificidade ou variância específica(ψ_i). Desta forma, têm-se as seguintes equações:

Equação da Variância da variável $\underline{X}_i$

$$V(X_i) = \ell_{i2}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 + \psi_i \qquad (3.1)$$

Equação das comunalidades

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + ...l_{im}^2 \tag{3.2}$$

Portanto as comunalidades são as maiores parcelas do total da variância de uma variável X_i. A segunda parcela é proveniente da variância específica de cada variável, representada por ψ_i.

Quanto mais a comunalidade se aproximar de 1, melhor será o modelo fatorial. Autores consideram boa comunalidade valores acima de 0,70.

3.1.2.5 Matriz dos Resíduos

A Análise Fatorial também permite o cálculo dos valores residuais ou matriz dos resíduos. Uma vez que a matriz de covariância S pode ser escrita como $S = \hat{L}\hat{L}' + \hat{\psi}$ tem-se uma estimativa da matriz de resíduos dada por:

$$MR = S - (\hat{L}\hat{L}' + \hat{\psi}) \tag{3.3}$$

onde:

MR é a matriz dos resíduos;

S é a matriz de covariância amostral;

$\hat{L}$ é matriz dos pesos estimados;

$\hat{L}'$ é a transposta da matriz dos pesos estimados;

$\hat{\psi}$ é a matriz das variâncias específicas estimadas.

A matriz dos resíduos é outra forma de se avaliar se o modelo fatorial está próximo da realidade, pois expressa a diferença entre as correlações e o produto dos pesos estimados e sua transposta, acrescidos dos erros. Valores próximos de zero indicam que o modelo Fatorial é adequado.

3.1.2.6 Rotação dos Fatores

Outra técnica muito interessante na análise fatorial é a Rotação dos Fatores. Tal técnica é empregada para otimizar os pesos fatoriais. Como o próprio nome diz, a rotação consiste em girar os eixos em um ângulo θ, oferecendo uma nova estrutura para os pesos de tal forma que cada variável tenha peso alto em um único fator e pesos mais baixos ou médios nos demais fatores. A rotação não produz uma estrutura visível quando m>2, ou seja, o número de fatores é maior que dois e quando isso ocorre programas computacionais são utilizados para executar a rotação e proceder aos cálculos dos fatores rotacionados, embora a visualização continue impossível.

A Rotação pode ser realizada no sentido horário(3.4) e anti-horários(3.5) e produzir a matriz T [18] ortogonal, que multiplicada pelos pesos estimados originais produzirá os novos pesos(3.6) conforme as equações matriciais que seguem:

$$T = \begin{bmatrix} \cos(\theta) & \text{sen}(\theta) \\ -\text{sen}(\theta) & \cos(\theta) \end{bmatrix} \text{ (rotação no sentido horário)} \qquad (3.4)$$

$$T = \begin{bmatrix} \cos(\theta) & -\text{sen}(\theta) \\ \text{sen}(\theta) & \cos(\theta) \end{bmatrix} \text{ (rotação no sentido anti-horário)} \qquad (3.5)$$

$$\hat{L}^{*} = \hat{L}T, \qquad (3.6)$$

pois $\sum = \hat{L}\hat{L}'+\hat{\Psi} = \hat{L}\hat{T}\hat{T}'\hat{L}'+\hat{\Psi} = \hat{L}\hat{L}'+\hat{\Psi}$

onde

$\hat{L}$ é a matriz dos pesos estimados originais;

$\hat{L}^{*}$ é a nova matriz dos pesos estimados rotacionada e

T é a matriz de transformação ortogonal.

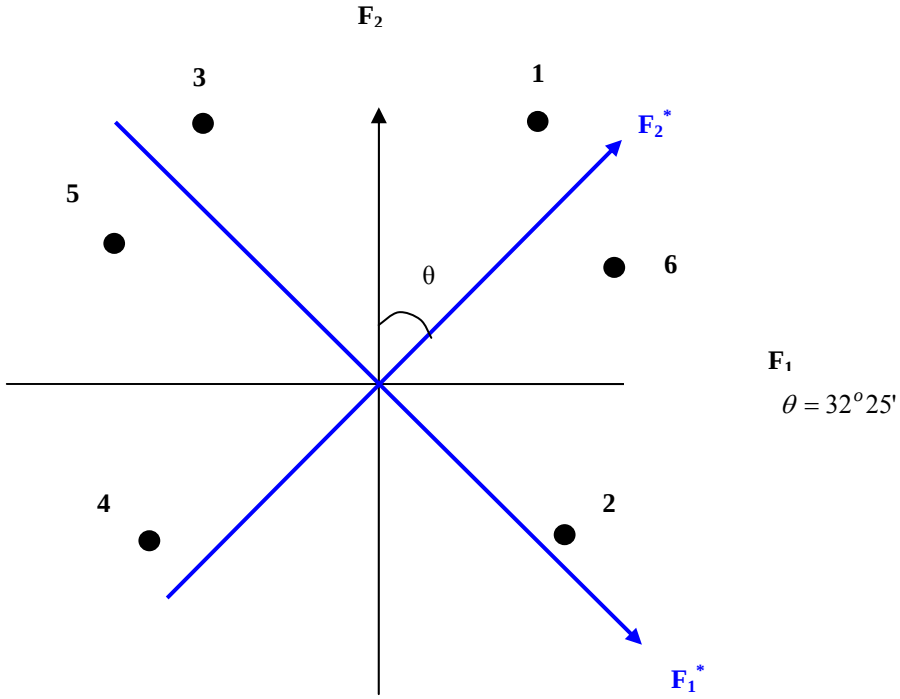
A técnica de rotação mais utilizada foi proposta por Kaiser e recebeu o nome de Rotação Varimax ou Normal Varimax.

---

[18] Exemplo tomado para dois fatores.

Segue um exemplo, no gráfico 3.1, da rotação de fatores para dois fatores variáveis.

GRÁFICO 3.1- EXEMPLO DE ROTAÇÃO APLICADA EM SEIS VARIÁVEIS COM DOIS FATORES.



FONTE: AUTOR

3.2 REDES NEURAIS ARTIFICIAIS.

> "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).

3.2.1 O Neurônio Biológico

O cérebro humano é considerado o mais fascinante processador baseado em carbono existente, sendo composto por aproximadamente 100 bilhões de neurônios[19].Todas as funções e movimentos do organismo estão relacionados ao funcionamento destas pequenas

células. Os neurônios estão conectados uns aos outros através de sinapses, e juntos formam uma grande rede, chamada REDE NEURAL. As sinapses transmitem estímulos através de diferentes concentrações de Na+ (Sódio) e K+ (Potássio), e o resultado disto pode ser estendido por todo o corpo humano. Esta grande rede proporciona uma fabulosa capacidade de processamento e armazenamento de informação.

A célula nervosa, ou neurônio foi identificado anatomicamente e descrito com detalhes, pelo neurologista espanhol Ramón y Cajal, no século XIX (1894). O neurônio e delimitado por uma fina membrana celular, como qualquer outra célula, que além da função biológica normal, possui outras propriedades que são fundamentais para o funcionamento elétrico da célula.

As manifestações elétricas de neurônios biológicos foram observadas pela primeira vez no século 19 por DuBois Reymond, com auxílio de Galvanômetros[20]. Porém somente em 1924, utilizando tubo de raios catódicos[21], foram efetivamente confirmadas por Erlanger e Gasser que foram laureados com o Nobel de Medicina (fisiologia) em 1944.

Nas duas décadas seguintes, 30 e 40, após trabalho de muitos pesquisadores, passou-se a entender o neurônio biológico como sendo basicamente um dispositivo computacional

Nos neurônios a comunicação é realizada através de impulso.Quando um impulso é recebido, o neurônio o processa, e passado um limite de ação, dispara um segundo impulso que produz uma substância neuro-transmissora o qual flui do corpo celular para o axônio (que por sua vez pode ou não estar conectado a um dendrito de outra célula). O neurônio que transmite o pulso pode controlar a freqüência de pulsos aumentando ou diminuindo a polaridade na membrana pós sináptica. Eles têm um papel essencial na determinação do funcionamento, comportamento e do raciocínio do ser humano. Ao contrário das redes neurais artificiais, redes neurais naturais não transmitem sinais negativos, sua ativação é medida pela freqüência com que emite pulsos, freqüência esta de pulsos contínuos e positivos. As redes naturais não são uniformes como as redes artificiais, e apresentam uniformidade apenas em alguns pontos do organismo. Seus pulsos não são síncronos ou assíncronos, devido ao fato de

---

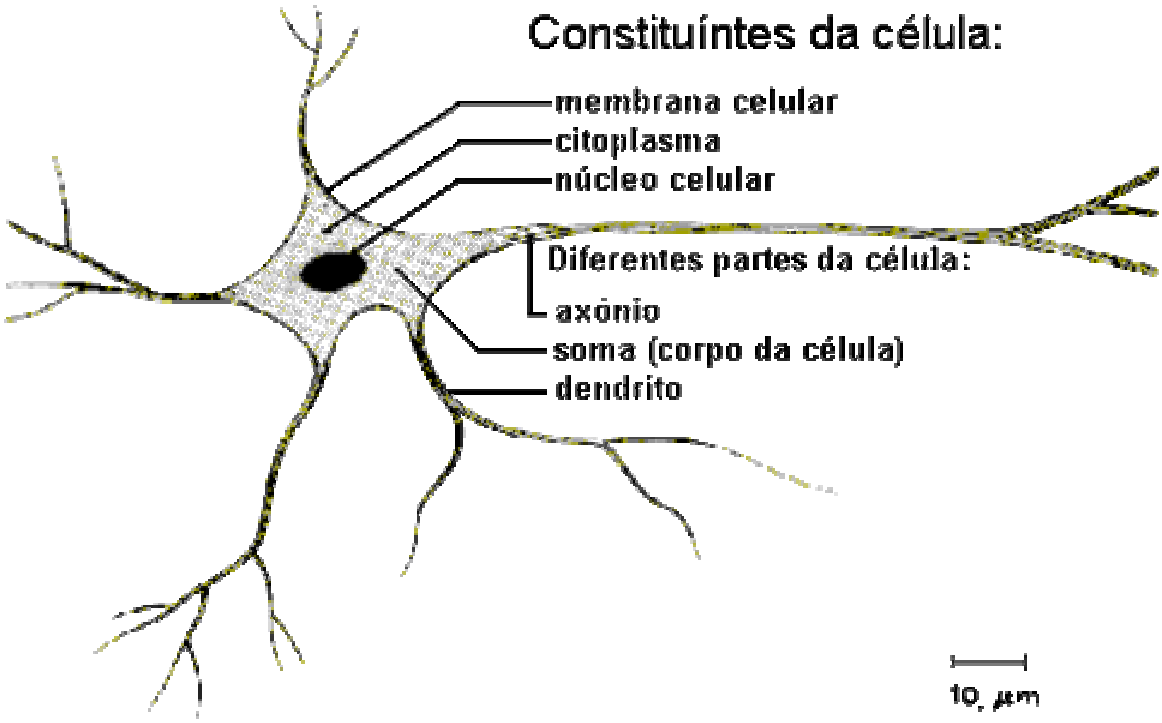
[19] Dado disponível no site:< http://numerologia.avancada.nom.br/funcionamento_do_cerebro.htm>

[20] Galvanômetros: aparelhos que registram pequenas correntes elétricas através de efeitos magnéticos.

[21] Tubo de Raios Catódicos: criação de Chookes-final do século XIX.

não serem contínuos, o que a difere de redes artificiais O neurônio divide-se nas seguintes partes, conforme indica a figura 3.2

FIGURA 3.2 – ESQUEMA DA CÉLULA NEURAL BIOLÓGIGA



FONTE: INTERNET[22]

- Os dendritos, que tem por função receber os estímulos transmitidos pelos outros neurônios;
- o corpo de neurônio, também chamado de soma, que é responsável por coletar e combinar informações vindas de outros neurônios;
- e o axônio, que é constituído de uma fibra tubular que pode alcançar até alguns metros, e é responsável por transmitir os estímulos para outras células.

3.2.2 Modelo matemático

O primeiro modelo matemático para o neurônio biológico foi proposto por

[22] Disponível no site www.din.uem.br/ia/neurais

McCulloch e Pitts em 1943. Neste modelo o neurônio recebe sinais oriundos de várias entradas e a combinação desses sinais é propagada ao corpo celular (soma).À medida que o tempo passa as entradas passam a ter uma maior ou menor influência sobre o processo que ocorre no corpo celular, ou seja, elas podem excitar ou inibir processamento. Esse comportamento é representado através do somatório ponderado por pesos das entradas, pesos esses que geralmente oscilam entre –1 e 1. Sendo assim, chama-se os valores dos sinais de entradas de $x_1, x_2, x_3, ...x_n$ e os pesos ponderados de $w_1, w_2, w_3, ...w_n$, temos que a entrada total no corpo celular dada pela seguinte fórmula:

$$ent_{total} = \sum_{i=1}^{n} x_i . w_i . \tag{3.7}$$

Considera-se também, em muitos casos um valor inicial para a entrada, denominado de *bias*, que possui valor unitário também ponderado por seu respectivo peso w_0, modificando a entrada que fica com o seguinte formato:

$$ent_{total} = \sum_{i=1}^{n} x_i . w_i + w_0 \tag{3.8}$$

ou ainda:

$$ent_{total} = \sum_{i=0}^{n} x_i . w_i . \tag{3.9}$$

Depois disso o valor total da entrada sofre transformação através da aplicação de uma função de ativação gerando o valor de saída. Que fica matematicamente representada por:

$$saída = f(ent_{total}) . \tag{3.10}$$

A função de ativação pode ser uma função matemática qualquer, porém geralmente são utilizadas funções lineares(reta) , funções degrau, função de Siebert[23], e mais recentemente estão sendo utilizadas funções sigmóides ou logística (logit) e também função tangente hiperbólica pois são simétricas e possuem as derivadas contínuas, são crescentes e monotônicas e possuem saída-resposta no intervalo $[-1,1]$.

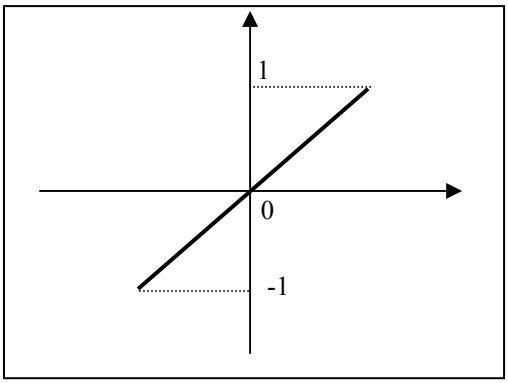
Seguem os gráficos representativos de algumas funções:

---

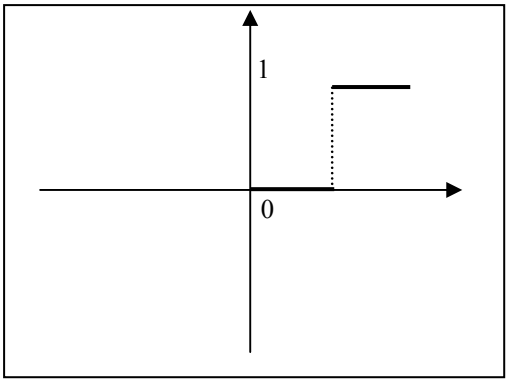
[23] Siebert: propôs uma função que recebeu seu nome, quando modelava neurônios no sistema auditivo dos vertebrados.

GRAFICO 3.2 – FUNÇÕES DE ATIVAÇÃO

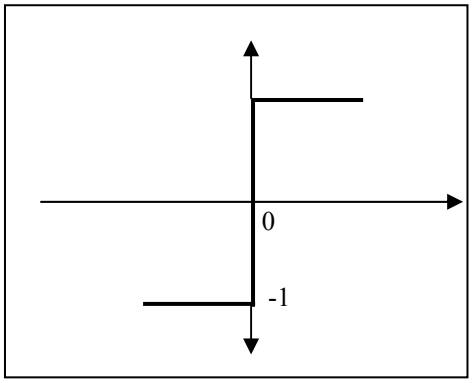
Função linear (reta)



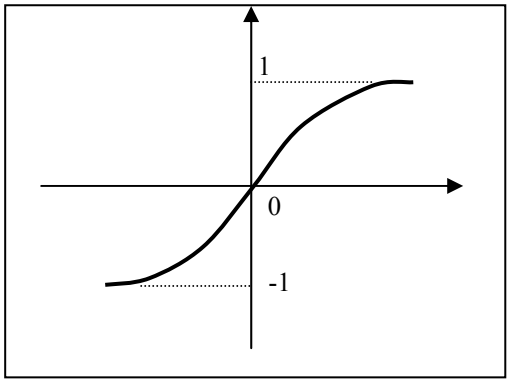
Função degrau I



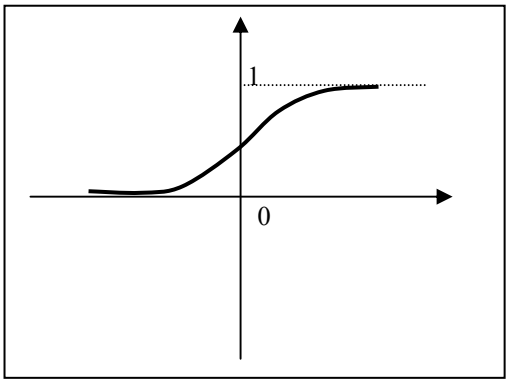
Função sinal II



Função tangente hiperbólica



Função logística ou sigmóide



3.2.3 Conceito e funcionamento de uma Rede Neural

As redes neurais artificiais consistem em um método de solucionar problemas de inteligência artificial, construindo um sistema que tenha circuitos que simulem o cérebro humano, inclusive seu comportamento, ou seja, aprendendo, errando e fazendo descobertas. São técnicas computacionais que apresentam um modelo inspirado na estrutura neural de

organismos inteligentes e que adquirem conhecimento através da experiência. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento, enquanto que o cérebro de um mamífero pode ter muitos bilhões de neurônios.

Um grafo direcionado é um objeto geométrico que consiste de um conjunto de pontos, chamados nós, ao longo de um conjunto de segmentos de linhas direcionadas entre eles. Uma rede neural é uma estrutura de processamento de informação distribuída paralelamente na forma de um grafo direcionado, com algumas restrições e definições próprias.

Os nós deste grafo são chamados elementos de processamento. Suas arestas são conexões, que funcionam como caminhos de condução instantânea de sinais em uma única direção, de forma que seus elementos de processamento podem receber qualquer número de conexões de entrada. Estas estruturas podem possuir memória local, e também possuir qualquer número de conexões de saída desde que os sinais nestas conexões sejam os mesmos. Portanto, estes elementos, têm na verdade uma única conexão de saída, que pode dividir-se em cópias para formar múltiplas conexões, sendo que todos carregam o mesmo sinal.

Então, a única entrada permitida para a função de transferência (que cada elemento de processamento possui) são os valores armazenados na memória local do elemento de processamento e os valores atuais dos sinais de entrada nas conexões recebidas pelo elemento de processamento. Os únicos valores de saída permitidos a partir da função de transferência são valores armazenados na memória local do elemento de processamento, e o sinal de saída do mesmo.

A função de transferência pode operar continuamente ou episodicamente. Sendo que no segundo caso, deve existir uma entrada chamada "*activate*" que causa o ativamento da função de transferência com o sinal de entrada corrente e com valores da memória local, e produzir um sinal de saída atualizado (ocasionalmente alterando valores da memória). E no primeiro caso, os elementos estão sempre ativados, e a entrada "*activate*" chega através de uma conexão de um elemento de processamento agendado que também é parte da rede.

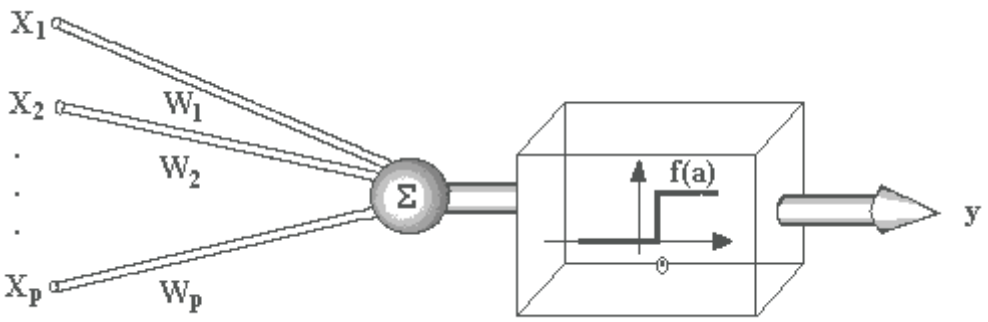
De forma geral, a operação de uma célula da rede se resume em:

- Sinais são apresentados à entrada;
- Cada sinal é multiplicado por um peso que indica sua influência na saída da unidade;
- É feita a soma ponderada dos sinais que produz um nível de atividade;
- Se este nível excede um limite, a unidade produz uma saída;

3.2.4 Histórico e acontecimentos na evolução das Redes Neurais Artificiais

As primeiras informações mencionadas sobre a neuro computação datam de 1943, em artigos de McCulloch e Pitts, em que sugeriam a construção de uma máquina baseada ou inspirada no cérebro humano, simulando o comportamento do neurônio natural, onde o neurônio possuía apenas uma saída, que era uma função de entrada *(threshold*) da soma do valor de suas diversas entradas conforme esquematizado na figura 3.3.

FIGURA 3.3 – NEURÔNIO ARTIFICIAL PROPOSTO POR MACCULLOCH E PITTS



FONTE: INTERNET[24]

Em 1949 Donald Hebb escreveu um livro intitulado "*The Organization of Behavior*" (A Organização do Comportamento) que perseguia a idéia de que o condicionamento psicológico clássico está presente em qualquer parte dos animais pelo fato de que esta é uma propriedade de neurônios individuais. Suas idéias não eram completamente novas, mas Hebb foi o primeiro a propor uma lei de aprendizagem especifica para as sinapses dos neurônios.

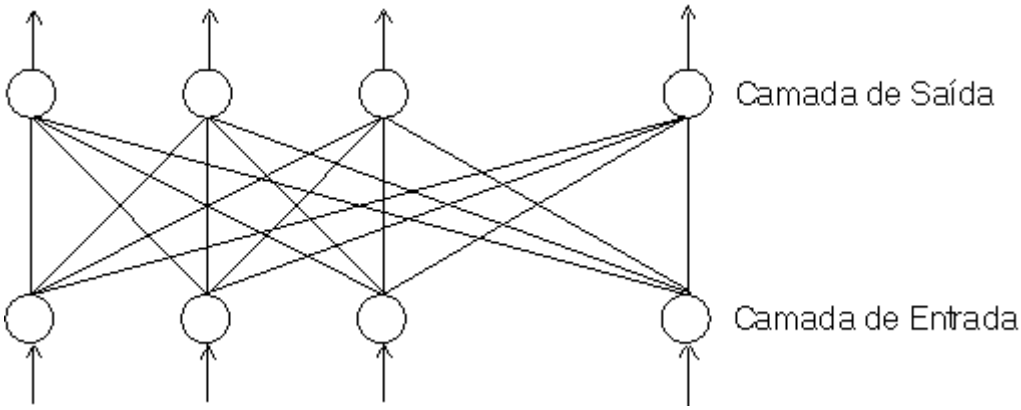
Também proveniente deste período foi a construção do primeiro neuro computador, denominado *Snark*, por Mavin Minsky, em 1951. O *Snark* operava com sucesso a partir de um ponto de partida técnico, ajustando seus pesos automaticamente, entretanto, ele nunca executou qualquer função de processamento de informação interessante, mas serviu de inspiração para as idéias de estruturas que o sucederam.

O primeiro neuro computador a obter sucesso (Mark I Perceptron) surgiu em 1957

[24] Disponível no site www.din.uem.br/ia/neurais

e 1958, criado por Frank Rosenblatt, Charles Wightman e outros. Devido à profundidade de seus estudos, suas contribuições técnicas e de sua maneira moderna de pensar, muitos o vêem como o fundador da neuro computação na forma em que a temos hoje. Seu interesse inicial para a criação do *Perceptron* era o reconhecimento de padrões.

FIGURA 3.4 – REDE DE PERCEPTONS PROPOSTA POR ROSENBLATT



FONTE: INTERNET[25]

Após Rosenblatt, Bernard Widrow, com a ajuda de alguns estudantes, desenvolveram um novo tipo de elemento de processamento de redes neurais chamado de *Adaline*, figura 3.5, equipado com uma poderosa lei de aprendizado, que diferente do *Perceptron* ainda permanece em uso.

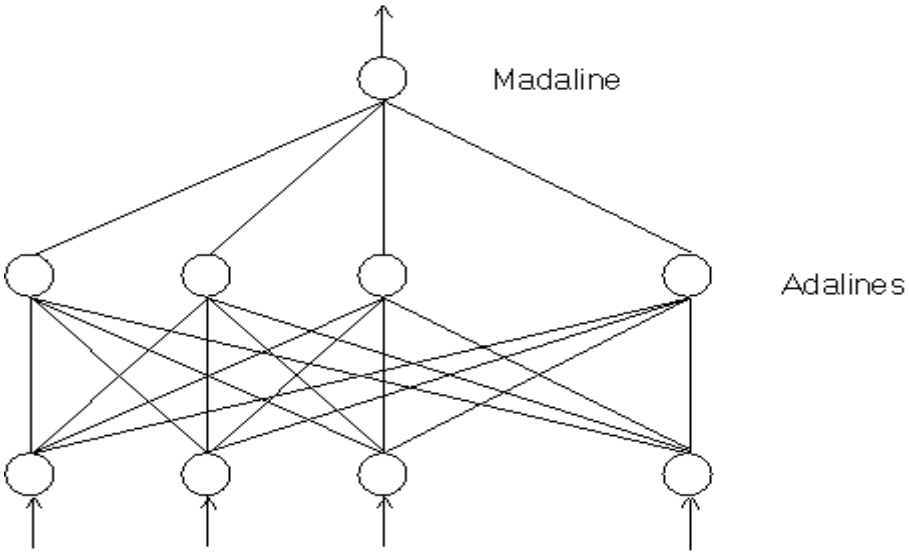
Um período de pesquisa silenciosa seguiu-se durante 1967 a 1982, quando poucas pesquisas foram publicadas devido aos fatos ocorridos anteriormente. Entretanto, aqueles que pesquisavam nesta época, e todos os que se seguiram no decorrer de treze anos conseguiram novamente estabelecer um campo concreto para o renascimento da área.

Nos anos 80, muitos dos pesquisadores foram bastante corajosos e passaram a publicar diversas propostas para a exploração de desenvolvimento de redes neurais bem como suas aplicações. Porém talvez o fato mais importante deste período tenha ocorrido quando Ira Skurnick, um administrador de programas da DARPA (*Defense Advanced Research Projects Agency*) decidiu ouvir os argumentos da neuro computação e seus projetistas, e divergindo dos caminhos tradicionais dos conhecimentos convencionais, fundou em 1983 pesquisas em neuro computação. Este ato não só abriu as portas para a neuro computação, como também

[25] Idem anterior

deu a DARPA o status de uma das líderes mundiais em se tratando de "moda" tecnológica.

FIGURA 3.5 – REDE ADALINE PROPOSTA POR WIDROW E OUTROS



FONTE: INTERNET[26]

.

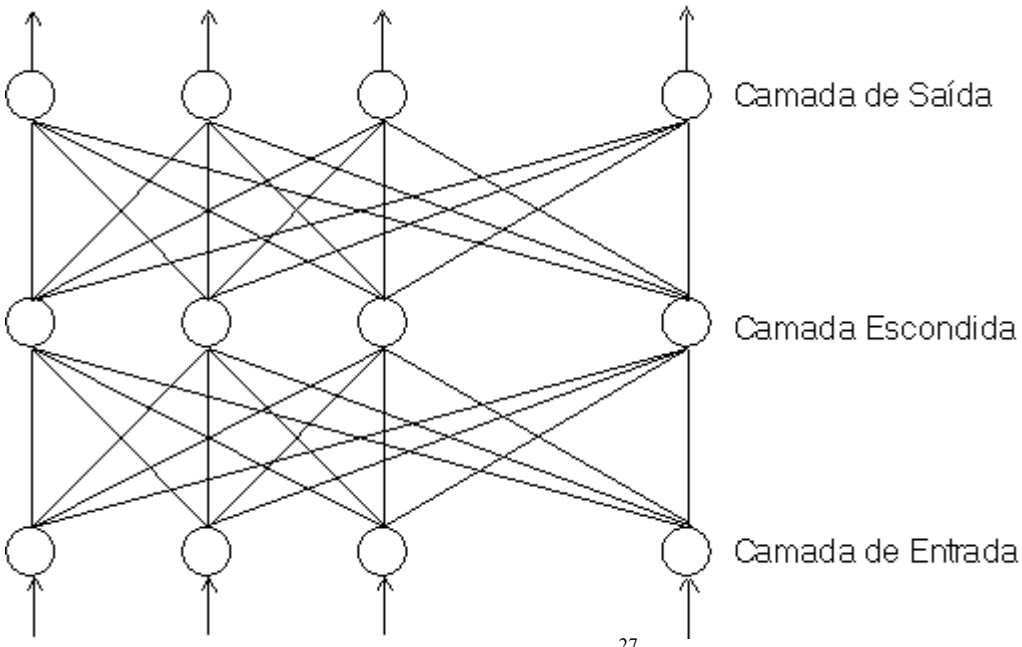
Outra "potência" que emergiu neste período foi John Hopfield, renomado físico de reputação mundial, se interessou pela neuro computação, e escreveu artigos que percorreram o mundo todo persuadindo centenas de cientistas, matemáticos, e tecnólogos altamente qualificados a se unirem esta nova área emergente.

Apesar de um terço dos pesquisadores da área terem aderido à mesma pela influência de Hopfield, foi em 1986 que este campo de pesquisa "explodiu" com a publicação do livro "*Parallel Distributed Processing*" (Processamento Distribuído Paralelo) editado por David Rumelhart e James McClelland.

Rumelhart, juntamente com Hinton e Williams introduziram o poderoso método *Backpropagation*, figura 3.6, algoritmo de retro-propagação que funciona da seguinte forma: A entrada é apresentada e propagada para frente através da rede, calculando as ativações para cada unidade de saída. Cada unidade de saída compara com o valor desejado, resultando em um valor de erro. Depois de calculados os erros em cada unidade são realizadas alterações nos pesos que é a etapa de retorno do algoritmo.

FIGURA 3.6 – ESTRUTURA DE REDE UTILIZANDO O ALGORITMO BACKPROPAGATION

[26] Idem anterior



FONTE: INTERNET[27]

Em 1987 ocorreu em São Francisco a primeira conferência de redes neurais em tempos modernos, a IEEE, *International Conference on Neural Networks*, e também foi formada a *International Neural Networks Society* (INNS). A partir destes acontecimentos decorreram a fundação do INNS *journal* em 1989, seguido do *Neural Computation* e do IEEE *Transactions on Neural Networks* em 1990.

Hoje em dia a quantidade de material a respeito de Redes Neurais cresceu acentuadamente. Com o advento do computador nos bancos escolares e da Internet, trabalhos, publicações e teses são inseridos todos os dias nos *sites* de Universidades do mundo todo. Novos algoritmos surgiram, modernas estruturas de redes também, porém todos tendo como alicerce os trabalhos e conceitos desenvolvidos por várias décadas.

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* indicaram que os comandos utilizados trabalham com o algoritmo *Backpropagation*, acrescidos de outro algoritmo, denominado *Levenberg-Marquadt* que é descrito na seqüência.

3.2.5 Rede *Backpropagation* com o algoritmo de Levenberg-Marquadt

[27] Idem anterior.

Enquanto o *backpropagation* padrão utiliza a descida de gradiente como método de aproximação do mínimo da função erro, o algoritmo de Levenberg-Marquardt (LM) utiliza uma aproximação pelo método de Newton. Esta aproximação é obtida a partir da modificação do método de Gauss-Newton.

O método de Gauss-Newton é aplicável a uma função de custo que é expressa como a soma de erros quadrados como indica a fórmula 3.11.

$$\varepsilon(\mathbf{w}) = \frac{\mathbf{1}}{\mathbf{2}} \sum_{i=1}^{n} e^2(i) \tag{3.11}$$

onde o fator de escala ½ é incluído para simplificar a análise subseqüente. Todos os termos de erro nesta fórmula são calculados com base no vetor de peso w que é fixo dentro de todo o intervalo de observação $1 \leq i \leq n$.

O sinal de erro *e(i)* é uma função do vetor de peso ajustável w. Dado um ponto de operação w (*n*), lineariza-se a dependência de *e(i)*em relação a w escrevendo-se matricialmente;

$$\mathrm{e}'(n,\mathrm{w}) = \mathrm{e}(n) + \mathrm{J}(n)\,(\mathrm{w} - \mathrm{w}(n)), \tag{3.12}$$

onde e (*n*) é o vetor de erro, dado por

$$\mathrm{e}(n) = [e(1), e(2),..., e(n)]^T \tag{3.13}$$

e J(*n*) é *a matriz jacobiana n*-por-*m* de e(*n*):

$$\mathbf{J}(n) = \begin{bmatrix} \frac{\partial e(1)}{\partial w_1} & \frac{\partial e(1)}{\partial w_2} & \cdots & \frac{\partial e(1)}{\partial w_m} \\ \frac{\partial e(2)}{\partial w_1} & \frac{\partial e(2)}{\partial w_2} & \cdots & \frac{\partial e(2)}{\partial w_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial e(n)}{\partial w_1} & \frac{\partial e(n)}{\partial w_2} & \cdots & \frac{\partial e(n)}{\partial w_m} \end{bmatrix}_{\mathbf{w}=\mathbf{w}(n)} \tag{3.14}$$

A jacobiana J(n) é a transposta da matriz de gradiente m-por-n ∇ e(n), onde

$$\nabla \text{ e}(n) = [\nabla \text{ e}(1), \nabla \text{ e}(2), ..., \nabla \text{ e}(n)] \tag{3.15}$$

o vetor de peso atualizado w(n+1) é assim definido por

$$\text{w}(n+1) = \arg_{\text{w}} \min\left\{\frac{1}{2}\left\|\text{e}'(n, \text{w})\right\|^2\right\} \tag{3.16}$$

Usando a equação 3.12 para calcular a norma euclidiana quadrática de e′ (n,w), obtem-se

$$\begin{aligned}\frac{1}{2}\left\|\mathbf{e}'(n, \mathbf{w})\right\|^2 &= \frac{1}{2}\left\|\mathbf{e}(n)\right\|^2 + \mathbf{e}^T(n)\mathbf{J}(n)(\mathbf{w} - \mathbf{w}(n)) \\ &\quad + \frac{1}{2}(\mathbf{w} - \mathbf{w}(n))^T \mathbf{J}^T(n)\mathbf{J}(n)(\mathbf{w} - \mathbf{w}(n))\end{aligned} \tag{3.17}$$

Assim, diferenciando-se esta expressão em relação à w e igualando-se o resultado a zero, obtem-se

$$\text{J}^T(n)\text{e}(n) + \text{J}^T(n)\text{J}(n)(\text{w} - \text{w}(n)) = 0 \tag{3.18}$$

Resolvendo-se esta equação para w, pode-se então escrever a partir da equação (3.16) vem

$$\text{w}(n + 1) = \text{w}(n) - (\text{J}^T(n)\ \text{J}(n)^{-1}\ \text{J}^T(n)\text{e}(n) \tag{3.19}$$

que descreve a forma pura do método de Gauss-Newton.

Diferentemente do método de Newton, que requer o conhecimento da matriz hessiana da função custo $\varepsilon(n)$, o método de Gauss-Newton requer apenas a matriz jacobiana do vetor de erro e(n). Entretanto, para que a iteração de Gauss-Newton seja computável, a matriz do produto $\text{J}^T(n)\text{J}(n)$ deve ser não singular.

Com relação a este último ponto, reconhece-se que $J^T(n)J(n)$ é sempre definida não negativamente. Para assegurar-se que ela seja não-singular, a jacobiana $J(n)$ deve ter *posto n*, em relação às linhas; isto é, as n linhas de $J(n)$ na matriz jacobiana 3.14 devem ser linearmente independentes (L. I). Infelizmente, não há garantia de que esta condição seja sempre satisfeita. Para resguardar-se, contra a possibilidade de que $J(n)$ seja deficiente em posto, a prática habitual é adicionar a matriz diagonal δI à matriz $J^T(n)J(n)$. O parâmetro δ é uma constante positiva pequena escolhida para assegurar que

$$J^T(n)J(n) + \delta I, \text{ seja definida positivamente para todo } n$$

Baseando-se nisto, o método de Gauss-Newton é implementado na forma ligeiramente modificada:

$$w(n + 1) = w(n) - (J^T(n)\, J(n) + \delta I)^{-1}\, J^T(n)\, e(n) \tag{3.20}$$

O efeito desta modificação é reduzido progressivamente à medida que o número de iterações, n, é aumentado. Nota-se também que a equação recursiva 4.20 é a solução da função de custo *modificada*:

$$\varepsilon(\mathbf{w}) = \frac{1}{2}\left\{\delta\|\mathbf{w} - \mathbf{w}(0)\|^2 + \sum_{i=1}^{n} e^2(i)\right\} \tag{3.21}$$

onde $w(0)$ é o *valor inicial* do vetor de peso $w(i)$. (HAYKIN, 2001).

3.3 ANÁLISE DE DISCRIMINANTE

Uma técnica estatística apropriada para discriminação e classificação é a análise discriminante (Hair et all., 1998; Johnson et all. 1998).

O objetivo principal desta técnica é classificar as observações em uma ou mais classes pré-determinadas, portanto trata-se de uma técnica de predição.

O princípio fundamental é obter uma regra que possa ser utilizada para classificar de forma otimizada uma nova observação em uma classe previamente identificada.

O vetor $\underline{X}$ 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 Linear 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.

3.3.1 Modelo Matemático

A concepção de Fischer é buscar por uma regra, sensível de Fischer. Para encontrar a função primeiramente deve-se pensar em transformar as observações multivariadas $\underline{X}$ nas observações univariadas, tal que os Y's nas populações π_1 *e* π_2 estejam separados o máximo possível. Para conseguir tal feito trabalha-se com combinações lineares de $\underline{X}$ que geram os Y's. Considerando que as médias dos Y's obtidos dos $\underline{X}$'s são dadas por:

μ_{1y} = média dos Y's obtidos dos $\underline{X}$'s pertencentes a população π_1 e

μ_{2y} = média dos Y's obtidos dos $\underline{X}$'s pertencentes a população π_2.

Então Fischer selecionou a combinação linear que maximiza o quociente entre a distância quadrática de μ_{1y} a μ_{2y} e a variabilidade (variância) dos Y's . Admitindo-se:

$$\mu_1 = E(\underline{X} \mid \pi_1): \tag{3.22}$$

valor esperado de uma observação multivariada de π_1

$$\mu_2 = E(\underline{X} \mid \pi_2): \tag{3.23}$$

valor esperado de uma observação multivariada de π_2

$$\Sigma = E[(\underline{X} - \underline{\mu_i})(\underline{X} - \underline{\mu_i})'], \quad \text{i} = 1,2 \tag{3.24}$$

Como a matriz de covariância $\underline{X}$ é:

$Y = \underline{C}'\underline{X}$

Então, considerando a combinação linear tem-se que:

$$\mu_{1Y} = E(Y \mid \pi_1) = E(\underline{C}'\underline{X} \mid \pi_1) = \underline{C}'E(\underline{X} \mid \pi_1) = \underline{C}'\,\underline{\mu_1} \tag{3.25}$$

$$\mu_{2Y} = E(Y \mid \pi_2) = E(\underline{C}'\underline{X} \mid \pi_2) = \underline{C}'E(\underline{X} \mid \pi_2) = \underline{C}'\,\underline{\mu_2} \tag{3.26}$$

e a variância é:

$$\sigma_Y^2 = V(Y) = V(\underline{C}'\underline{X}) = \underline{C}'V(\underline{X})\underline{C} = \underline{C}'\,\Sigma\underline{C} \tag{3.27}$$

que é a mesma para ambas as populações

Fischer afirmou que: "a melhor combinação linear vem da razão entre o quadrado da distância entre as médias e a variância de Y", desta forma tem-se:

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_y^2} = \frac{(\underline{C}'\underline{\mu_1} - \underline{C}'\underline{\mu_2})^2}{\underline{C}\Sigma\underline{C}} = \frac{(\underline{C}'\underline{\delta})^2}{\underline{C}'\Sigma\underline{C}} \tag{3.28}$$

que é a razão a ser maximizada, onde:

$\underline{\delta} = \underline{\mu_1} - \underline{\mu_2}$ e $V(\underline{X}) = \Sigma$;

Maximiza-se a razão por:

$$\underline{C} = K\Sigma^{-1}\underline{\delta} = K\Sigma^{-1}(\underline{\mu_1} - \underline{\mu_2}) \tag{3.29}$$

para qualquer K $\neq 0$, então por simplicidade adota-se K=1 e tem-se:

$$Y = \underline{C}'\underline{X} = (\underline{\mu_1} - \underline{\mu_2})'\Sigma^{-1}\underline{X} \tag{3.30}$$

que é a chamada Função Discriminante Linear de Fischer.

A F.D.L de Fischer transforma as populações π_1 e π_2 multivariadas em populações univariadas de tal forma que as médias das populações univariadas estejam separadas o máximo possível considerando ainda que as variâncias populacionais sejam as mesmas para ambas as populações

Quando um novo indivíduo surge dentro do problema estudado e deseja-se pré-dizer onde deve ser alocado, utiliza-se a F.D.L de Fischer onde $\underline{X}_0$ é o vetor com as observações do novo indivíduo. Desta forma tem-se:

$$Y_0 = (\underline{\mu_1} - \underline{\mu_2})'\Sigma^{-1}\underline{X_0} \tag{3.31}$$

como sendo o valor calculado da F.D.L de Fischer para o novo indivíduo. Para verificar onde ele deve ser alocado considera-se o ponto médio entre as populações univariadas como m. Então:

$$m = \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) = \frac{1}{2}(\underline{C}'\underline{\mu_1} + \underline{C}'\underline{\mu_2}) = \frac{1}{2}\underline{C}'(\underline{\mu_1} + \underline{\mu_2}) = \frac{1}{2}[(\underline{\mu_1} - \underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1} + \underline{\mu_2})] \tag{3.32}$$

assim, tendo-se calculado os valores de Y_0 e m pode-se classificar o novo indivíduo através

da seguinte regra:

Se $Y_0 \geq \mathrm{m}$, aloca-se X_0 em π_1 e se $Y_0 < m$, aloca-se X_0 em π_2.

3.3.2 Probabilidade de erro no uso da Análise de Discriminante.

Uma melhor visualização do erro pode ser dada com o cálculo da taxa aparente de erro e apresentação da matriz de confusão que é a comparação entre acertos e erros na classificação dos novos elementos provenientes das populações π_1 e π_2. A matriz de confusão em sua forma genérica segue no quadro 3.1, e permite o cálculo da Taxa Aparente de erro, como segue.

QUADRO 3.1 – MODELO DE MATRIZ DE CONFUSÃO

		Classificação Prevista	
		π_1	π_2
Classificação Real	π_1	n_{11}	n_{12}
	π_2	n_{21}	n_{22}

FONTE: AUTOR

Onde:

n_1 número de itens total em π_1

n_2 número de itens total em π_2

n_{11} número de itens de π_1 classificados corretamente como de π_1.

n_{12} número de itens de π_2 classificados incorretamente como de π_1.

n_{22} número de itens de π_2 classificados corretamente como de π_2.

n_{12} número de itens de π_1 classificados incorretamente como π_2.

Cálculo da Taxa aparente de erro:

$$APER = \frac{n_{12} + n_{21}}{n_1 + n_2}, \tag{3.33}$$

que é a proporção das observações classificadas incorretamente.

Ainda é possível calcular a probabilidade de erro para classificações equivocadas, ou seja, alocar-se em π_1 quando na realidade o indivíduo pertence a π_2 e vice-versa, usando a aproximação gaussiana. Esse cálculo é feito mediante a transformação das médias univariadas em medidas padronizadas, e utiliza-se a distribuição normal de probabilidade, para verificar a probabilidade de erro, através da área sob a curva normal.

Desta forma tem-se:

$$z_i = \frac{m - \mu_{iY}}{V(Y)} \tag{3.34}$$

para, i= 1 e 2 onde:

m: ponto médio entre as populações univariadas;

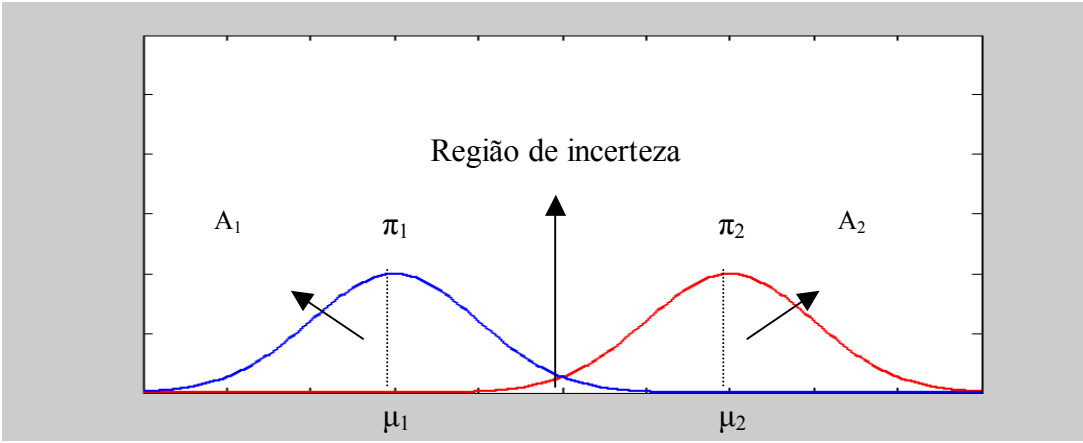
μ_{iY} : médias univariadas das população π_1 e π_2;

V(Y): variância das populações univariadas (valor igual para as duas pop.)

z_i : médias padronizadas das populações univariadas.

Graficamente tem-se:

GRAFICO 3.3 – EXEMPLO DE DISTRIBUIÇÕES NORMAIS PADRONIZADAS DAS POPULAÇÕES π_1 π_2.



$A_1 = P(Y_0 > m \mid \underline{X}_0 \in \pi_1)$ $A_2 = P(Y_0 \leq m \mid \underline{X}_0 \in \pi_2)$

FONTE: AUTOR - MATLAB

A análise do gráficos da distribuição normal para as duas amostras π_1 e π_2 na região de confluência das duas curvas, indica a área ou percentual de erro na classificação de um novo indivíduo. Esta área é a única região de dúvida ou incerteza para a Função Discriminante de Fischer.

3.4 REGRESSÃO LOGÍSTICA

Sempre que se deseja estudar as relações entre um conjunto de variáveis independentes com uma variável dependente se esta considerando um problema multivariado. Na análise de tal problema, geralmente usamos algum modelo matemático para lidar com as complexas inter-relações entre as diversas variáveis.

A regressão logística é uma 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.

3.4.1. A Função Logística

O modelo de regressão logística é também conhecido como modelo logístico e é baseado na função sigmóide, $f(z)$, dada por :

$$f(z) = \frac{1}{1+e^{-z}} \tag{3.35}$$

ou ainda

$$f(z) = \frac{e^{z}}{1+e^{z}} \quad \text{(habitualmente mais usada)} \tag{3.36}$$

Portanto, pode-se observar que a função logística varia entre 0 e 1, ($0 \leq f(z) \leq 1$) e essa é a principal razão do modelo logístico ser usado para descrever uma dicotômica probabilidade de algo acontecer ou não acontecer,ou seja, variável dicotômica.

Existem ainda outras duas características de *f*(*z*) que tornam o modelo logístico largamente utilizado:

- *z* representa um índice que combina a contribuição de diversos fatores de risco, e *f*(*z*) representa o risco (probabilidade) de que um evento ocorra, para um dado *z*;
- possui forma de "S", indicando que o efeito de *z* em *f*(*z*) é mínimo até que algum "gatilho" seja disparado, depois aumenta rapidamente até que algum nível seja alcançado, voltando a crescer lentamente.

3.4.2 O Modelo Matemático Logístico

A partir da função logística, $f(z)$, pode-se obter o modelo logístico, escrevendo z como a soma linear das variáveis independentes e substituindo na função vem:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{3.37}$$

Assim a função logística passa a ser representada como a contribuição das variáveis independentes X e os parâmetros desconhecidos $\beta_0, \beta_1, ...\beta_n$. Para que ela torne-se o modelo logístico deve-se escrever na forma específica , substituindo $f(z)$, por $\pi(x)$ como segue:

$$E(Y \mid x) = \pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_n x_n}} \tag{3.38}$$

onde $E(Y \mid x)$ é a esperança de Y dado $\underline{x}$.

3.4.3 Transformação Logit

A Transformação Logit ou logito é interessante e necessária, pois estabelece uma relação linear entre as variáveis explicativas e a transformada da variável resposta. A demonstração da transformação segue abaixo

Partindo-se da equação 3.38

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_n x_n}}$$

chamando $g = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_n x_n$ (3.39)

e resolvendo para g(x), tem-se:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] \tag{3.40}$$

sendo que a função logit $g(x)$ é agora linear nos seus parâmetros.

3.4.4 Ajuste do Modelo de Regressão Logística

O ajuste pode ser realizado através do método de mínimos quadrados não linear ou

pelo método da máxima verossimilhança (Hosmer e Lemeshow, 1989). O método de máxima verossimilhança é sempre mais indicado por possuir propriedades ótimas. Desta forma tem-se as estimativas dos parâmetros $\beta_0, \beta_1, ... \beta_n$ dadas por $\hat{\beta}_0, \hat{\beta}_1, ... \hat{\beta}_n$ que por sua vez alimentarão o modelo estimado e a logit estimada que ficam representadas na seguinte forma:

- Modelo estimado, $\hat{\pi}(x) = \dfrac{e^{\hat{\beta}_0 + \hat{\beta}_{11} x_1 + ... + \hat{\beta}_n x_n}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n x_n}}$

- Logit Estimada , $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n x_n$.

Assim, após estimados os valores dos coeficientes $\hat{\beta}_0, \hat{\beta}_1, ... \hat{\beta}_n$, basta substituí-los nas funções acima para encontrar o valor estimado de $\hat{\pi}(x)$, que logicamente será um valor entre 0 e 1, ou seja, $0 \leq \hat{\pi}(x) \leq 1$. De acordo com esse valor classifica-se o novo indivíduo em sua respectiva categoria dicotômica.

3.4.5 Verificação do Ajuste.

A qualidade do modelo ajustado através da regressão logística pode ser verificada através da comparação entre os valores observados e os valores preditos para a variável resposta. De forma bastante simples, constrói-se uma tabela com os valores observados versus os valores preditos e verifica-se a existência de uma relação linear entre essa duas quantidade conforme segue no quadro 3.2, elaborado para um exemplo de 100 indivíduos classificados em portadores de uma virose (1) e não portadores (0).

QUADRO 3.2 EXEMPLO DE CLASSIFICAÇÃO DOS CASOS PARA 100 INDIVÍDUOS

Variável Dicotômica	**Preditas em 0**	**Preditas em 1**	**% de acertos**
0	45	12	78,94%
1	14	29	67,44%

FONTE: AUTOR

CAPÍTULO 4

4 METODOLOGIA E RESULTADOS OBTIDOS

Neste capítulo, as técnicas escolhidas e comentadas no capítulo III, serão utilizadas na resolução do problema apresentado. Para que as técnicas de *Data Mining* sejam efetivamente utilizadas, é fundamental que se faça alguns comentários a respeito da origem do banco de dados e as adequações que foram realizadas nesse banco para que pudessem ser utilizados nos algoritmos escolhidos.

4.1 BANCO DE DADOS

Como foi citado no capítulo II, as duas primeiras fases do processo KDD envolvem a elaboração do banco de dados ideal para realização das tarefas de *Data Mining*. O primeiro passo para realizar o processo consiste em obter a compreensão do domínio da aplicação e as metas do usuário final. Já o segundo passo é a criação de um conjunto de dados meta, selecionando um conjunto de dados, ou estabelecendo um subconjunto de variáveis ou exemplos de dados, sobre os quais a descoberta será executada.

Após análise dos bancos de dados disponíveis nas empresas varejistas pesquisadas, concluiu-se que esse banco de dados era muito pobre em informações a respeito dos funcionários, pois continha somente algumas variáveis, tais como: nome do funcionário, idade, sexo, endereço e grau de instrução. Para realização das tarefas de predição em *Data Mining* se fez necessário à elaboração de um banco de dados mais robusto, com o maior número de informações possíveis e que possam efetivamente ser usadas na aplicação das técnicas.

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

4.2 O QUESTIONÁRIO

A elaboração de um questionário para coleta de dados não é tarefa simples. A primeira dificuldade que surge é o que se deve perguntar e quantas perguntas devem-se fazer. Outras dúvidas, dizem respeito ao formato do questionário, ou seja, como as perguntas serão apresentadas e de que forma o questionado poderá respondê-las de forma rápida e objetiva. Como o objetivo principal do trabalho é predizer se um determinado funcionário estará ou não satisfeito com o seu emprego as perguntas foram divididas em dois grupos: as perguntas relacionadas com a satisfação ou insatisfação dos funcionários e as perguntas 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.

Outro fator importante considerado é o tempo necessário para responder às perguntas. Um questionário muito extenso poderia levar o pesquisado ao desânimo e as informações não teriam a veracidade necessária. Portanto optou-se por elabora 30 perguntas no total, dispostas em somente duas folhas de papel no formato A4, contento 10 perguntas relacionadas com satisfação ou insatisfação e mais 20 perguntas relacionadas às características pessoais.

Depois de concluído, o questionário foi aplicado em 14 lojas de uma mesma rede de supermercados num total de 826 funcionários.

Segue as 30 perguntas realizadas e o questionário completo faz parte do apêndice n^{o}. 1

Perguntas relacionadas às características pessoais num total de 20

- Pergunta 1: Idade
- Pergunta 2: Sexo
- Pergunta 3: Estado Civil
- Pergunta 4: Número de Filhos
- Pergunta 5: Em sua família você é o filho nr.....
- Pergunta 6: Grau de instrução
- Pergunta 7: Religião
- Pergunta 8:Mês de nascimento
- Pergunta 9: Animal preferido

- Pergunta 10: Cor preferida
- Pergunta 11: Seu tempo de folga é ocupado com...
- Pergunta 12: Setor em que trabalha
- Pergunta 13: Função em que trabalha
- Pergunta 14: Tempo de trabalho na empresa
- Pergunta 15: Distância da residência ao trabalho
- Pergunta 16: Meio de condução ao trabalho
- Pergunta 17: Número de registros em carteira
- Pergunta 23: Faixa salarial
- Pergunta 26: Tempo de permanência que julga ideal
- Pergunta 27: Existe motivo que levaria a sua rescisão.

Perguntas relacionadas à satisfação do funcionário no emprego num total de 10.

- Pergunta 18: Quanto a sua função você está...
- Pergunta 19: Quanto à loja em que você trabalha você está...
- Pergunta 20: Seu relacionamento com os colegas de trabalho é...
- Pergunta 21: Seu relacionamento com superiores diretos é...
- Pergunta 22: Seu relacionamento com a direção da empresa é...
- Pergunta 24: Seu nível salarial em comparação a outros é...
- Pergunta 25: Você trocaria de empresa pelo mesmo salário.
- Pergunta 28: De maneira geral seu trabalho é...
- Pergunta 29: De maneira geral a empresa é...
- Pergunta 30: A oferta de emprego nos próximos 5 anos vai melhorar...

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*. Essa fase, como foi citado no capítulo II, é a etapa que requer mais tempo dentro do processo KDD. Os dados precisam ser limpos, filtrados e reduzidos para que se obtenha o conjunto de dados que efetivamente possam ser utilizados e recebam a denominação de Data Warehouse.

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 de resposta até 16 pontos, responderam entre ótimo e bom (1 e 2), portanto estão satisfeitos. Funcionários cujo somatório ficou entre 17 e 26 responderam entre bom e regular (3 e 4), portanto estão parcialmente satisfeitos. Finalmente funcionários cujo somatório apresentou valor superior a 26 pontos tiveram concentração de respostas entre, ruim e péssimo (5 e 6) e foram classificados como insatisfeitos.

Segue a escala adotada na classificação do somatório:

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

Posteriormente os funcionários cujo somatório apresentaram respostas 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 4.1 mostra uma parte deste somatório, a escala de resultados e a classificação binária final.

QUADRO 4.1 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:AUTOR

Resolvido o problema da classificação binária, o próximo passo concentrou-se com a parte do questionário que envolve 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 segue no quadro 4.2.

QUADRO 4.2 – RESPOSTAS DAS 20 PERGUNTAS DE 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: AUTOR

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 o questionário com as respostas foi analisado e algumas modificações foram realizadas, algumas variáveis foram acrescidas, outras retiradas, na tentativa de criar uma segunda matriz que foi denominada de matriz preparada. Segue a descrição das perguntas que foram retiradas, acrescidas e modificadas, sendo que, as demais perguntas e respectivas respostas ficaram no formato do questionário original (apêndice 1).

- Pergunta nº 3: Estado Civil. Foram criados apenas quatro grupos como segue:
  - Solteiro, identificado com 1
  - Casado, identificado com 2
  - Amasiado, identificado com 3
  - Outro: identificado com 4
- Pergunta nºs. 12 e13: As perguntas número 12 e 13, função e setor, foram unidas em uma só caracterizando somente o setor onde o funcionário trabalha e reduzindo o número de setores (que conforme os anexos 1 e 2 eram muitos), para somente 3 setores: administrativo, atendimento e logística[28]. Ficando como segue:
  - Setor de atendimento ao público, denominado com 1
  - Setor de logística, denominado com 2
  - Setor administrativo, denominado com 3.
- Pergunta nº 14: Tempo de trabalho: Não foi modificada apenas levado em consideração que se o novo candidato ao emprego, como ainda não tem tempo de casa, receberia o número 0(zero) para esse quesito. Portanto as opções continuaram as mesmas do questionário inicial com respostas variando de 0 a 10.
- Pergunta nº 26: Tempo que você julga necessário que um trabalhador fique na mesma empresa: Essa pergunta foi excluída, pois a maior parte dos entrevistados não soube responder.
- Pergunta nº 31: Esta pergunta não existia anteriormente e foi acrescida a matriz preparada, trata-se do ramo de negócio característico de cada loja. Quatro ramos foram identificados e numerados de 1 a 4.
- Pergunta nº32: Outra variável acrescentada à matriz, diz respeito à localização da loja em cada cidade sede, levando-se em conta a população e faturamento

das lojas. Desta forma 5 possibilidades foram identificados e numerados de 1 a 5.

- Pergunta nº33: última variável acrescida à matriz, trata-se do IDH-INDICE DE DESENVOLVIMENTO HUMANO[29], de cada município. Esse índice por ser um número decimal, foi multiplicado por 100, para transformá-lo em inteiro e evitar que a matriz ficasse mal condicionada.

Após todas as modificações citadas terem sido concluídas a nova matriz modificada, ficou no formato 826 x 21, ou seja, 826 indivíduos com 21 componentes (características), conforme o quadro 4.3.

QUADRO 4.3 – RESPOSTAS DAS 20 PERGUNTAS DE CARACTERÍSTICAS PESSOAIS - MATRIZ MODIFICADA.

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: AUTOR

A matriz modifica, apresentada no quadro 4.3 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 foi a redução de dimensionalidade, verificando as variáveis que possuem menor influência na explicação da resposta de predição.

---

[28] Os três setores foram criados com base no histórico do R.H da empresa pesquisada.

[29] Fonte: O IDH dos municípios foram retirados do Censo do IBGE-2000.

4.3 REDUÇÃO DE DIMENSIONALIDADE

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

4.3.1. Autovalores e Variância Explicada

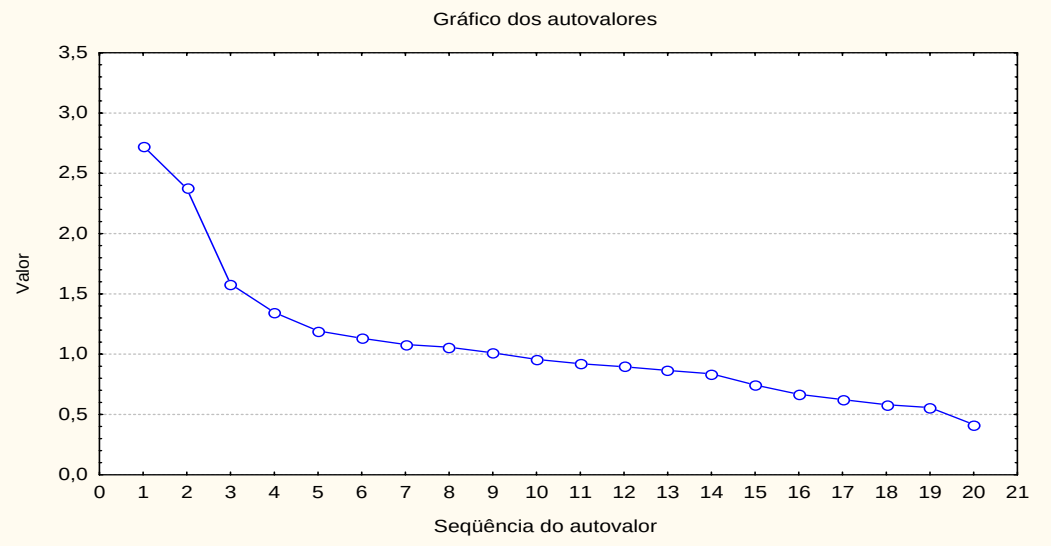
Um dos cálculos realizados com o auxilio do *Statistica* foi o de encontrar os autovetores através da Análise das Componentes Principais, conforme indica o quadro 4.4.

QUADRO 4.4 – AUTOVALORES E % DE VARIÂNCIA EXPLICADAA

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

FONTE: AUTOR

GRÁFICO 4.1 – AUTOVALORES DAS 21 VARIÁVEIS ORIGINAIS



Esse quadro indica a presença de 14 autovalores maiores que 0,8[30] .A última coluna da tabela reforça a idéia que 14 variáveis são as principais, dentre as 21 iniciais, pois explicam aproximadamente 84% da variância total. O gráfico 4.1 mostra os autovalores das 21 variáveis, proporcionando melhor visualização dos resultados. As variáveis de menor importância foram identificadas analisando os autovetores, são elas: Idh, idade, condução, setor, tipo de loja, estado civil e nr. de registros em carteira. A tabela completa com os resultados dos 21 autovalores segue no apêndice 2.

4.3.2 PESOS FATORIAIS

Outro dado de suma importância que o pacote de análise fatorial do *Statistica* oferece é o cálculo dos pesos fatoriais que permite distinguir quais fatores entre os 14 estão carregados[31] com quais variáveis. Para cálculo dos pesos fatoriais optou-se por uma rotação dos fatores, também conhecida como Rotação Varimax. Conforme descrito no capítulo 3, a rotação permite que os fatores sejam translacionados próximos de variáveis que os carregam com mais intensidade. O quadro 4.5 mostra os pesos fatoriais, já rotacionados, com destaque para as variáveis mais carregadas. A análise dessa tabela mostra claramente o carregamento do fator número 1 pelas variáveis 1, 3 e 16, o segundo fator está carregado pelas variáveis 19,20 e 21, enquanto que a variável 17 (número de registros em carteira), não carrega nenhum fator com peso acima de 0,7.

Também foram calculadas as comunalidades (apêndice 03), a matriz dos resíduos (apêndice 04), que conforme citado no capítulo 3 também são dados utilizados para certificação de que o modelo fatorial está próximo da realidade e a matriz dos escores fatoriais (quadro 4.6) que será utilizada para reduzir a dimensão de um novo indivíduo.

---

[30] Kaiser considera razoável, autovalores superiores a 1,0 como sendo os principais.

[31] O carregamento em uma variável é indicado pelo peso fatorial. Pesos maiores que 0,7 indicam bom carregamento, ou seja, o fator específico está sendo explicado por essa variável.

QUADRO 4.5 – PESOS FATORIAIS ROTACIONADOS

	Fatores													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
VAR1	**0,7839**	0,1041	0,1344	0,0350	-0,0406	0,1562	-0,2183	-0,0501	-0,0247	0,0209	0,0924	-0,0185	0,0621	0,1065
VAR2	0,1290	0,0285	**-0,7164**	0,0517	0,0241	0,1486	-0,0055	-0,1434	-0,0882	-0,1168	0,1817	0,0695	0,0299	-0,3725
VAR3	**0,7785**	0,0654	-0,1075	0,0120	0,0321	-0,0320	0,0230	-0,0627	0,0154	0,0458	-0,0357	0,0250	-0,0776	-0,0376
VAR4	0,0348	0,0756	0,0322	0,0128	0,0008	0,0108	0,0222	0,0018	0,0186	0,0213	-0,0071	**-0,9843**	-0,0147	0,0131
VAR5	0,0660	0,0142	0,0020	0,0382	0,0179	**0,9773**	-0,0250	0,0122	0,0095	0,0310	0,0060	-0,0111	-0,0462	0,0766
VAR6	-0,2919	-0,0303	0,0229	-0,0206	0,0278	-0,1152	0,1075	0,1060	0,0266	0,0021	-0,0882	0,0158	0,0181	**-0,8185**
VAR7	-0,0351	-0,0641	0,0335	0,0243	0,0410	0,0456	0,0048	0,0144	-0,0129	0,0120	0,0137	-0,0148	**-0,9868**	0,0149
VAR8	0,0089	-0,0208	0,0149	0,0312	-0,0119	0,0089	-0,0074	-0,0195	**0,9955**	-0,0138	-0,0151	-0,0184	0,0128	-0,0134
VAR9	0,0151	-0,0903	0,0361	0,0061	**0,9764**	0,0177	0,0140	-0,0086	-0,0126	-0,0063	0,0319	-0,0024	-0,0412	-0,0271
VAR10	0,0324	0,0136	-0,0205	-0,0237	0,0302	0,0059	0,0043	0,0540	-0,0144	0,0070	**0,9742**	0,0063	-0,0139	0,0531
VAR11	0,0826	0,0378	0,0590	0,0157	-0,0066	0,0310	-0,0423	-0,0040	-0,0142	**0,9885**	0,0065	-0,0214	-0,0120	0,0065
VAR12	0,0987	0,1464	**0,7759**	0,0511	0,0366	0,0579	0,0512	-0,0069	-0,0252	-0,0023	0,0459	0,0215	-0,0105	-0,0915
VAR13	0,0652	0,0787	0,0153	0,0228	-0,0126	0,0214	**-0,9482**	0,0090	0,0085	0,0416	-0,0083	0,0248	0,0044	0,0701
VAR14	0,0243	-0,0821	-0,0191	**0,8638**	0,0314	0,0080	0,1036	0,0472	0,0025	0,0279	-0,1014	0,0888	-0,0532	0,1153
VAR15	0,0452	-0,1457	0,0986	**0,8218**	-0,0290	0,0385	-0,1514	-0,0690	0,0338	-0,0121	0,0835	-0,1134	0,0297	-0,1290
VAR16	**0,7348**	-0,1204	0,1553	0,0284	0,0136	-0,0103	0,0623	0,0994	0,0135	0,0266	0,0061	-0,0569	0,0822	0,1875
VAR17	0,2973	0,0081	0,5999	0,1302	0,0262	0,0703	-0,2773	-0,1304	-0,0192	0,0004	0,0791	-0,0559	-0,0170	-0,3977
VAR18	0,0149	0,0444	-0,0244	0,0128	0,0065	-0,0117	0,0040	**-0,9684**	0,0199	0,0041	-0,0537	0,0016	0,0143	0,0629
VAR19	-0,0477	**-0,7208**	-0,2216	-0,0863	0,2121	-0,0047	-0,0679	0,1246	0,0239	-0,0122	-0,0595	0,1053	0,0069	-0,0109
VAR20	-0,0191	**0,8207**	-0,0802	-0,2413	0,0905	0,0106	-0,0992	0,0986	-0,0041	0,0149	-0,0373	0,0562	0,0513	-0,0232
VAR21	-0,0268	**-0,9494**	-0,0479	0,1027	0,0250	-0,0064	0,0638	0,0427	0,0018	-0,0203	-0,0020	0,0508	-0,0333	-0,0348
Expl.Var	1,9859	2,2037	1,6058	1,5273	1,0186	1,0302	1,0976	1,0408	1,0050	0,9994	1,0348	1,0211	1,0042	1,0739
Prp.Totl	0,0946	0,1049	0,0765	0,0727	0,0485	0,0491	0,0523	0,0496	0,0479	0,0476	0,0493	0,0486	0,0478	0,0511

FONTE: AUTOR (STATISTICA)

4.3.3 ESCORES FATORIAIS

Finalizando a aplicação da Análise Fatorial foram calculados os escores fatoriais, após a Rotação Varimax. São esses escores que substituem a matriz modificada (quadro 4.3) no formato 826 X 21, por uma nova matriz, reduzida para 826 X 14, ou seja, 826 indivíduos e 14 escores fatoriais. Portanto os escores fatoriais são os novos dados e sendo assim a tarefa de redução de dimensionalidade do problema foi concluída.

Parte dessa nova matriz segue no quadro 4.6

QUADRO 4.6 ESCORES FATORIAS ROTACIONADOS

Escores Fatorias														
Rotação Varimax normalizada														
Extração de Componentes Principais														
	Fatores													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	-1,3393	1,1053	1,2624	0,746	-0,5248	-0,9036	0,1718	-1,3775	0,9362	0,6959	-0,6558	-0,0008	0,2691	0,6396
2	-1,2806	0,9223	0,9424	-0,9211	-0,8148	2,1515	0,7298	1,0402	0,1913	-0,2918	-0,9237	0,677	0,3005	-0,1315
3	-0,8419	0,9749	-0,7704	-1,436	0,0619	0,8093	0,514	0,3648	0,4	0,7442	-0,6615	-0,3046	-0,4063	-0,8765
4	1,0379	1,1308	-0,9129	1,1364	0,2315	-1,1962	0,4903	-0,72	-1,294	-0,3382	0,1607	0,6506	0,2862	-0,3871
5	-1,3399	0,8323	0,8959	-0,95	-0,7233	0,065	0,3184	-1,3883	-1,2759	0,8278	-0,8823	0,6873	0,2675	0,0267
6	-1,4539	0,973	0,7825	0,3828	-0,6133	1,5302	0,4271	-1,2936	0,1136	-1,397	-0,8857	-1,1166	0,3403	0,7627
7	0,7746	1,0642	-0,6916	-0,8752	0,0817	-0,3924	-0,698	0,7727	1,2935	0,6496	1,4715	0,8029	0,0475	-0,3986
8	-0,3354	1,0161	-0,1283	0,6487	-0,6426	0,2075	0,3697	-1,1387	1,1276	0,8954	-0,5249	-0,181	0,179	0,0345
9	-0,6754	0,9843	-0,3125	0,2242	-0,6315	-1,2311	0,233	0,6058	0,5039	0,8136	0,7562	-3,0047	0,0968	0,423
10	0,5968	1,0382	-1,0335	0,0244	-0,7356	-1,1374	-0,3108	0,7676	1,0591	0,7194	0,934	-1,2129	-0,0818	-1,6292
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
821	0,4145	-2,0814	3,578	-0,1863	2,3514	1,4599	-1,2246	-1,7844	0,827	-1,6259	-0,568	0,511	0,7227	-3,5264
822	-1,141	-1,6413	0,1248	0,367	0,224	0,5981	0,0378	1,4608	1,6106	-1,2323	-0,362	0,5363	0,4417	0,7007
823	1,0109	-1,7567	1,3704	0,6998	0,3685	-1,0045	0,6067	1,2647	0,9268	-0,7615	0,0759	-0,955	0,7441	0,4019
824	-0,1978	-2,054	1,0644	0,4449	0,22	1,3765	-1,9981	-1,6044	-1,6199	0,7767	1,1588	-1,2542	0,4049	-1,3722
825	0,9627	-1,7286	0,7721	0,7781	-0,6528	0,3265	-0,0541	0,0012	-1,1956	0,5711	1,475	0,919	0,7023	-0,8068
826	0,2519	-1,6459	1,3678	0,7874	-0,5332	-0,9745	-0,2198	0,797	0,0588	0,7416	-0,9858	0,9366	0,3309	-0,5016

FONTE: AUTOR

Após a redução de dimensionalidade o passo seguinte é o de separar essa matriz em dois grandes grupos:

- o grupo dos indivíduos que farão parte do treinamento das técnicas e darão origem os pesos e *bias* para as redes neurais, aos coeficientes de Fischer na análise de discriminante e aos valores dos coeficientes estimados no modelo de Regressão Logística.
- o grupo dos indivíduos restantes que farão parte da verificação ou teste das técnicas.

4.4 TREINAMENTO E TESTE

Conforme se verificou em trabalhos anteriores é comum nesse estágio separar o grupo utilizando-se 70% dos indivíduos para treinamento, enquanto que os 30 % restantes são direcionados para a fase de teste.

Os grupos de treinamento e teste foram separados com aleatoriedade para evitar

vícios e acúmulos que podem deteriorar os resultados. Para garantir aleatoriedade foram gerados no *software Excel* 108 números aleatórios entre um e 359, e mais 140 números entre 360 e 826, pois os primeiros 359 indivíduos possuem variável resposta 0 (insatisfeitos) e os 467 restantes possuem resposta 1 (satisfeitos), correspondendo respectivamente a 30% do total de indivíduos.

Separadas as matrizes de treinamento e teste as técnicas de *Data Mining* foram empregadas e os resultados seguem nos próximos itens.

4.5 TÉCNICA REDES NEURAIS.

A implementação da técnica Redes Neurais foi feita utilizando algoritmo desenvolvido no *software MATLAB*, e, também utilizado ferramentas, oriundas do próprio pacote do *MATLAB*.

O algoritmo utiliza a matriz de treinamento (578 x 14), que corresponde a 70 % dos dados, e o vetor resposta (578 x 1) para executar a rede e buscar os pesos e as bias. Ao mesmo tempo, já se alimenta a rede também com a matriz de treinamento (248 x 14), que corresponde a 30% dos dados, e seu respectivo vetor resposta (248 x 1). Optou-se por uma rede *backpropagation* com correção de erro pelo algoritmo de Levenberg-Marquadt que faz parte do *default* da função utilizada no *MATLAB*. Ainda foi necessário informar para o *software* o número de neurônios na camada de entrada, na camada escondida e na saída, bem como a função de ativação, o número de iterações e a margem de erro desejada.

A técnica Redes Neurais foi rodada várias vezes, sempre variando a topologia da rede, ou seja, o número de neurônios na camada escondida ou ainda o número de iterações, pois cada vez que se roda a rede novos valores são calculados para os pesos e a bias.

Após várias tentativas, constatou-se que os melhores resultados aparecem quando se usa 8 ou 10 neurônios para a camada escondida, apesar dos testes realizados utilizaram também 6, 8, 10 e 12 neurônios. Outro fator notado na prática foi no que diz respeito ao número de iterações (épocas). O número ideal de iterações foi 50. Números maiores foram utilizados tais como 100, 200 e até 1000 iterações, porém quando passam de 50 iterações a convergência diminui acentuadamente e o tempo computacional aumenta consideravelmente sem provocar influência nos resultados.

Segue o quadro 4.7 com a demonstração de alguns resultados da técnica Redes Neurais, contendo o número de neurônios na camada escondida, o número de iterações os

percentuais de erro no treinamento e no teste e também a função de ativação usada.

QUADRO 4.7 – PERFORMANCE DE ALGUMAS REDES – ERRO PERCENTUAL NO TREINAMENTO

rede	treinamento	teste	neurônios	iterações	função
1	34,990%	56,620%	6	30	logit
2	42,310%	51,530%	6	50	logit
3	24,780%	68,720%	8	50	logit
4	28,062%	62,120%	8	50	logit
5	28,900%	62,290%	8	50	logit
6	33,730%	59,290%	8	50	logit
7	26,970%	65,540%	8	50	logit
8	36,340%	59,820%	8	50	logit
9	27,510%	62,30%	8	50	logit
10	28,290%	64,470%	8	50	logit
11	**40,910%**	**49,970%**	**8**	**50**	**logit**
12	42,850%	51,000%	8	50	logit
13	39,260%	51,200%	8	50	logit
14	37,160%	55,720%	8	50	logit
15	**23,110%**	**65,160%**	**8**	**100**	**logit**
16	23,220%	66,980%	12	50	logit
17	28,980%	66,650%	12	100	logit

FONTE: AUTOR

No quadro 4.7 pode-se destacar a rede número 15, na qual obteve-se o menor erro percentual para a fase de treinamento 23,11%. Destaca-se também que a rede número 11 na qual obteve-se o melhor resultado para a fase de teste e onde também os erros percentuais da fase de treinamento (40,91%) e teste (49,97%) ficaram os mais próximos possíveis uns dos outros. Sendo assim escolheu-se como modelo mais apropriado, a rede número 15. Na seqüência segue os resultados completos dessa rede mostrando os valores das bias pesos e performances.

4.5.1 Pesos e Bias

Os pesos e bias calculados pela rede neural são importantíssimos, pois são coeficientes de multiplicações e adições que serão utilizados juntamente com as funções logit, para predizer se um novo indivíduo, com suas características individuais, pertence ao grupo dos satisfeitos e insatisfeitos. Desta forma, seguem no quadro 4.8 os resultados alcançados.

QUADRO 4.8 RESULTADOS DA REDE NEURAL NR 15.

Pesos da camada intermediária: 8 neurônios							
w1= matriz 8 x 14							
-1,6266	42,0387	-5,1773	18,3617	8,0502	-15,8193	-22,6611	
11,3054	-12,6235	5,2005	-1,7447	-9,6829	18,5151	-5,6980	
-3,8545	-23,3745	-38,8943	3,4056	29,5585	-9,6679	0,3302	
-29,5652	4,3048	-40,6589	-4,3595	-8,6501	-22,2447	-18,1095	
16,0603	24,3330	19,8139	-36,2319	-2,2957	-16,9290	-0,9547	
-16,6568	-59,9423	-11,9750	1,9268	10,9344	27,5351	-8,0295	
-1,4738	-9,8628	-6,3047	-35,8058	-28,0389	25,9833	-25,6372	
19,0535	17,0797	6,8688	-0,8150	-29,0918	-2,3349	1,9159	
-9,2537	7,7462	1,0600	5,3894	37,8841	-1,3941	-31,7172	
-16,5652	-0,6133	-7,5281	0,6483	-16,9636	-6,9550	9,3606	
17,4696	7,1127	0,9107	15,4409	33,8238	-1,5587	-19,2830	
-55,0399	-2,0969	5,2413	-24,2788	-8,9841	-18,7038	28,0881	
-0,4411	21,4999	-29,5385	35,7550	-20,6594	-4,7001	-2,1642	
22,2519	22,2349	35,5512	24,5813	-17,3348	-3,1952	11,0107	
-0,9517	7,3871	26,6172	11,7808	-11,2190	31,7601	2,2803	
5,2599	50,2533	-35,0413	-33,1989	7,7691	-43,1246	-46,0401	
Pesos da camada de saída: 1 neurônio							
w2 = vetor 1 x 8							
2,2390	1,9370	-1,0981	0,7990	1,0282	-0,6375	-1,0324	-1,7513
Bias da camada intermediária e da saída							
b1 vetor 8 x 1	**b2**	-0,8579					
44,3649							
-7,9437							
-28,782							
-58,118							
-17,365							
12,1782							
-9,8028							
25,5332							

FONTE: AUTOR

4.6 TÉCNICA DA FUNÇÃO DISCRIMINANTE LINEAR DE FISHER

A implementação da técnica da análise discriminante linear de Fisher foi utilizada por meio do *software MATLAB*, com função previamente desenvolvida. A entrada de dados se faz com duas matrizes: o grupo π_1, dos insatisfeitos, denominada matriz X_1 e o grupo π_2, dos satisfeitos, denominada matriz X_2. Essas matrizes representam 70% do total dos dados,

ou seja, nesse ponto utilizou-se as matrizes de treinamento.

A função executa vários cálculos entre eles destacam-se: as médias dos dois grupos, as matrizes covariâncias dos grupos, a matriz estimada $S_{p,}$ os coeficientes da função discriminante, as médias univariadas das populações, as variâncias (que nesse caso são iguais), e também transforma as médias univariadas em medidas normalizadas propiciando a avaliação da probabilidade de erro de classificação equivocada, conforme descrito no capítulo 3.

Seguem na quadro 4.9 os coeficientes da função discriminante que formam o vetor que será usado na matriz de teste e na caracterização de novos indivíduos e os demais dados comentados acima.

QUADRO 4.9 – DADOS DA APLICAÇÃO DA FUNÇÃO DISCRIMINANTE LINEAR DE FISHER

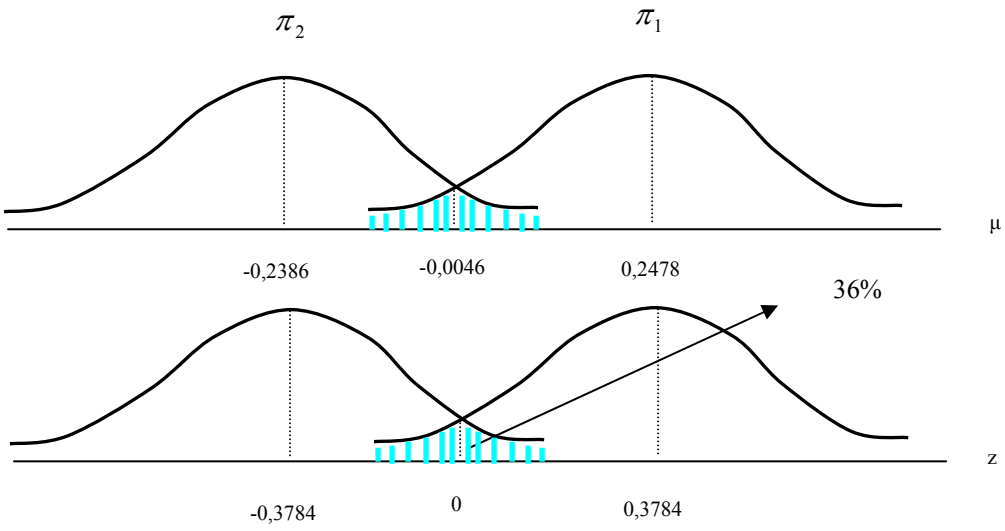
Coeficientes da função discriminante de Fisher													
-0,0893	-0,3514	-0,1631	-0,0150	-0,0048	0,0086	0,1680	0,5151	0,0811	0,1361	-0,1494	0,1148	0,0596	-0,0404
Média univariada				-0,046									
Média da população 1				0,2478									
Média da população 2				-0,2386									
variância da população 1				0,4864									
variância da população 2				0,4864									
Variância normalizada 1				-0,3784									
Variância normalizada 2				0,3487									

FONTE: AUTOR

4.6.1 Probabilidade de Erro

A probabilidade de erro, ou seja, classificar em π_1 quando o indivíduo pertence a π_2 e vice-versa é realizada utilizando-se as médias univariadas transformadas em variável z (normal) e a curva de distribuição normal acumulada padronizada. Consultando a tabela da distribuição normal, constatou-se que a área abaixo da curva com incertezas na classificação é de aproximadamente 36%, que representada graficamente fica conforme o gráfico 4.2.

GRAFICO 4.2 – DISTRIBUIÇÕES NORMAIS PADRONIZADAS DAS POPULAÇÕES π_1 π_2 .



FONTE: AUTOR

4.6.2 Teste dos Coeficientes de Fisher

O vetor dos coeficientes de Fischer multiplicado pala matriz de teste produz como resposta um número para cada indivíduo. Esse número , denominado como variável 'Y', está entre 0 e 1, ou seja $0 \leq Y \leq 1$. Comparando-se esse número com a média univariada das duas populações é possível pré-dizer se o indivíduo fará parte da população π_1 ou π_2 ou ainda se está a direita ou a esquerda da média univariada.

A multiplicação foi feita no *software Matlab* e apresentou os resultados constantes no quadro 4.10.

QUADRO 4.10 RESULTADOS PARCIAIS DA TÉCNICA DISCRIMINANTE DE FISCHER

nr	valor estimado	predição	realidade	resultado
1	0,545916809	0	0	correto
2	0,369512394	0	0	correto
3	-1,292562581	1	0	errado
4	-0,349378389	1	0	errado
5	-1,043670724	1	0	errado
6	0,957896427	0	0	correto
7	-0,571200401	1	0	errado
8	-0,181311238	1	0	errado
9	-0,044206878	1	0	errado
10	0,149198985	0	0	correto
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
97	1,119304455	0	0	correto
98	0,676910036	0	0	correto
99	1,340439225	0	0	correto
100	0,576356645	0	0	correto
101	1,101419527	0	0	correto
102	1,057373198	0	0	correto
103	0,777923909	0	0	correto
104	1,340828836	0	0	correto
105	0,457720842	0	0	correto
106	0,676694402	0	0	correto
107	1,100832908	0	0	correto
108	0,71834014	0	0	correto
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
245	-0,835130009	1	1	correto
246	1,492225043	0	1	errado
247	0,9146762	0	1	errado
248	-0,855374485	1	1	correto

FONTE: AUTOR

Os resultados completos podem ser observados na matriz de confusão indicada no quadro 4.11.

QUADRO 4.11 MATRIZ DE CONFUSÃO PARA A FUNÇÃO DISCRIMINANTE LINEAR DE FISCHER

		Classificação preditas	
	populações	Pi 1	Pi 2
	Pi 1	76	32
Classificação real	Pi 2	60	80

FONTE: AUTOR

4.7 TÉCNICA DA REGRESSÃO LOGÍSTICA

Na implementação da técnica Regressão Logística utilizou-se o *software Statistica,* pois além de apresentar recursos de cálculos e gráficos é de fácil manuseio e utiliza pouco tempo de processamento.

A análise dos dados foi feita alimentando o *software* com a matriz de treinamento, acrescentada da coluna das respostas, portanto a matriz a ser analisada ficou com formato 578 x 16.

O *software* possui um pacote especial na opção *Nonlinear Estimation,* que depois de alimentado com a matriz oferece a possibilidade de várias análises estatísticas, entre elas está a Regressão Logística (*Logistic Regression*). Para concretização da análise deve-se informar quais são as variáveis independentes e a variável dependente e ainda o método utilizado para estimativa dos parâmetros.

4.7.1 Estimativa dos Parâmetros

Os parâmetros ($\beta_0, \beta_1 ... \beta_{14}$) foram estimados pelo próprio pacote do *Statistica*, que para tanto se utiliza do método da máxima verossimilhança, num contexto de mínimos quadrados não lineares, aplicando-se o algoritmo de quase Newton, conforme comentado no capítulo 3.

Os resultado dos parâmetros estimados segue no quadro 4.12.

QUADRO 4.12 BETAS ESTIMADOS PELO MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Modelo:Regressão Logística														
Parâmetros estimados														
beta 0	beta 1	beta 2	beta 3	beta 4	beta 5	beta 6	beta 7	beta 8	beta 9	beta 10	beta 11	beta 12	beta 13	beta 14
0,763	0,093	0,342	0,177	0,020	0,003	-0,006	-0,211	-0,507	-0,081	-0,137	0,151	-0,118	-0,065	0,038

FONTE: AUTOR

4.7.2 Classificações Corretas

A matriz de confusão contendo os acertos e os erros na fase de treinamento é gerada no próprio pacote dos *Statistica* e segue no quadro 4.13

QUADRO 4.13 COMPARAÇÃO ENTRE CLASSIFICAÇOES CORRETAS E ERRADAS

Classificação dos casos				
	Pred.	Pred.		
	0	1	Percentual de acerto	
0	130,0000	121,0000	51,7928	%
1	87,0000	240,0000	73,3945	%

FONTE: AUTOR

4.7.3 Equações da Logit Estimada e do Modelo Estimado.

De posse dos parâmetros betas estimados na análise de regressão foi possível confeccionar as equações do modelo estimado e da logit estimada que podem ser agora utilizadas para realizar a fase de teste. As equações seguem abaixo.

Logit estimada:

$$\hat{g}(x) = 0.763 + 0.930x_1 + 0.342x_2 + 0.177x_3 + ... + 0..151x_{11} - 0.118x_{12} - 0.065x_{13} + 0.038x_{14} \qquad (4.1)$$

Modelo Estimado:

$$\hat{\pi}(x) = \frac{e^{0.763+0.930x_1+0.342x_2+0.177x_3+...+0..151x_{11}-0.118x_{12}-0.065x_{13}+0.038x_{14}}}{1+e^{0.763+0.930x_1+0.342x_2+0.177x_3+...+0..151x_{11}-0.118x_{12}-0.065x_{13}+0.038x_{14}}} \qquad (4.2)$$

Cada linha da matriz de teste com suas respectivas variáveis, transformadas previamente em escores fatoriais, (x_1,x_2...x_{14}) é substituída em 4.2 possibilitando o cálculo da variável binária que indicará se o indivíduo em estudo pertence a população π_1 ou. π_2.

4.7.4 Resultados da Regressão Logística

O quadro 4.14 indica os resultados parciais da matriz de teste utilizando os parâmetros estimados na regressão e a função logística. Apresenta também o resultado da função logit estimada e do modelo estimado.

QUADRO 4.14 – RESULTADOS PARCIAIS DA TÉCNICA DE REGRESSÃO LOGÍSTICA

nr	g(x)	Pi(x)	realidade	resultado
1	-0,292	0,428	0	correto
2	-0,135	0,466	0	correto
3	1,675	0,842	0	errado
4	0,658	0,659	0	errado
5	1,297	0,785	0	errado
6	-0,674	0,338	0	correto
7	0,847	0,700	0	errado
8	0,431	0,606	0	errado
9	0,283	0,570	0	errado
10	0,138	0,534	0	errado
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
97	-0,815	0,307	0	correto
98	-0,388	0,404	0	correto
99	-1,052	0,259	0	correto
100	-0,293	0,427	0	correto
101	-0,823	0,305	0	correto
102	-0,732	0,325	0	correto
103	-0,479	0,382	0	correto
104	-1,069	0,256	0	correto
105	-0,197	0,451	0	correto
106	-0,356	0,412	0	correto
107	-0,789	0,312	0	correto
108	-0,412	0,398	0	correto
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
245	1,222	0,772	1	correto
246	-1,192	0,233	1	errado
247	-0,614	0,351	1	errado
248	1,248	0,777	1	correto

FONTE: AUTOR

Os resultados finais da técnica de Regressão Logística podem ser visualizados na matriz de confusão para os indivíduos de teste, conforme o quadro 4.15.

QUADRO 4.15 – MATRIZ DE CONFUSÃO PARA A MATRIZ DE TESTE NA TÉCNICA REGRESSÃO LOGÍSTICA

	Predição			percentuais
	0	1		de acerto
0	62	46	108	57.4%
1	40	100	140	71.4%

O quadro 4.15 deixa evidente que o percentual de acerto para indivíduos satisfeitos é maior que os insatisfeitos, ratificando a matriz de confusão da fase de treinamento que mostrou percentuais muito próximos aos encontrados na fase de teste.

4.8 COMPARAÇÃO ENTRE AS TÉCNICAS

Os resultados coletados nas três técnicas de predição permitiram a comparação entre elas e a análise da melhor técnica.

Na fase de treinamento a técnica Redes Neurais mostrou-se mais apropriada, atingindo acertos de 76,89%, porém na fase de teste os resultados não foram satisfatórios.

A técnica Análise de Discriminante de Fischer mostrou-se mais apropriada para classificações de funcionários insatisfeitos, atingindo 70,37% na fase de teste.

A técnica de Regressão Logística apresentou resultados bastante próximos tanto para o treinamento, quanto para o teste, embora os melhores resultados tenham acontecido com os funcionários satisfeitos para fase de teste, com percentual de 71,40%. O quadro 4.16 demonstra os resultados obtidos, com destaque para os melhores resultados.

QUADRO 4.16 – COMPARATIVO DE ACERTO ENTRE AS TÉCNICAS

PERCENTUAIS DE CLASSIFICAÇÕES CORRETAS				
Técnica	Treinamento		Teste	
	satisfeitos	insatisfeitos	satisfeitos	insatisfeitos
Redes Neurais	**76,89%**		50,03%	
Discriminante de Fisher	64%		57,14%	**70,37%**
Regressão Logística	73,39%	51,79%	**71,40%**	57,40%

Desta forma a técnica mais indicada para predição foi a técnica Regressão Logística que proporcionou resultados mais próximos entre teste e treinamento, além de apresentar um bom percentual para indivíduos satisfeitos.

4.9 NOVOS INDIVÍDUOS

Após 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 no modelo logístico, pode-se agora fazer a predição de novos indivíduos, identificando em que grupo eles serão

enquadrados.

Para realizar a predição primeiramente deve-se aplicar o questionário para o novo indivíduo, 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 x 21 que deverá ser transposto e normalizado utilizando-se as médias e os desvios padrões de cada variáveis originais, previamente calculadas. 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:

$$z_i = \frac{x_i - \bar{x}_1}{s_i} \tag{4.3}$$

onde:

x_i é a i-ésima resposta da variável original.

$\bar{x}_1$ 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, (4.4) sofrerá agora a multiplicação pela matriz dos coeficientes dos escores fatoriais $(L'.L)^{-1}.L'$, gerada na análise fatorial (conforme o quadro 4.17), através da equação (4.5), produzindo o vetor f_i no formato (14 x 1) , ou seja, quatorze fatores para cada indivíduo conforme segue.

$$\underline{Z} = \begin{bmatrix} z_1 \\ z_2 \\ . \\ z_{21} \end{bmatrix} \tag{4.4}$$

$$\underline{f_i} = (L'L)^{-1}L'.\underline{Z} \tag{4.5}$$

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

$\underline{Z}$: é o vetor das variáveis padronizadas.

$\underline{f}$: é o novo vetor reduzido.

QUADRO 4.17 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: AUTOR

O quadro 4.18 mostra cinco novos candidatos ao emprego, com as 21 respostas do questionário, as médias e os desvios padrões das variáveis originais, a variáveis normalizadas e a redução para 14 fatores.

QUADRO 4.18 - CINCO NOVOS INDIVÍDUOS COM 14 FATORES

	novos indivíduos					média	d.p	novos ind. Normalizados					redução para 14 fatores				
variáveis	**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
animal	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					
registros	4	2	2	3	5	2.59	1.62	0.87	-0.37	-0.37	0.25	1.49					
salario	2	1	2	2	1	1.40	1.06	0.57	-0.38	0.57	0.57	-0.38					
rescisão	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:AUTOR

Esses novos vetores receberão os pesos e bias e a funções logit, os coeficientes de Fisher e os parâmetros betas da regressão e retornarão 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 desenvolvido planilha do Excel e parte desta planilha com as respostas segue no quadro 4.19.

QUADRO 4.19 RESULTADOS DAS PREDIÇÕES PARA CINCO NOVOS INDIVÍDUOS

novo indivíduo	redes	predição	fisher	predição	regressão	predição
y1	0,93054	satisfeito	1,5686	insatisfeito	0,908806	satisfeito
y2	0,33170	insatisfeito	0,1422	insatisfeito	0,706702	satisfeito
y3	0,66195	satisfeito	2,7306	insatisfeito	0,96805	satisfeito
y4	**0,57265**	**satisfeito**	**-0,1851**	**satisfeito**	**0,647769**	**satisfeito**
y5	0,37405	insatisfeito	0,4008	insatisfeito	0,768351	satisfeito

FONTE: AUTOR

Analisando-se o quadro 4.19 pode-se concluir, guardadas as devidas margens de erro, que o indivíduo que deve ser contratado é o ‘y4’, pois nas três técnicas o resultado predito para esse indivíduo foi a satisfação.

CAPÍTULO V

5 CONCLUSÃO

A tarefa de pré-dizer satisfação e insatisfação de funcionários, não são nada simplistas. 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 pré-dizer 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, profundo, de cada técnica e do processo de *Data Mining* 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 é a que requer maior tempo em todo o processo KDD.

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 de grupos previamente definidos 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, área da estatística 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 *Statistica*, sendo que nas técnicas Redes Neurais, Fischer e Análise Fatorial, trabalhou-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 pré-dizer, 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.

5.1 TRABALHOS FUTUROS

O 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, também citadas no capítulo 2.

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 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 utilizar a abordagem de Lachenbruch (MARQUES, 2004), como forma de avaliar a eficiência da regra de classificação elaborando um programa no próprio *Matlab*, que procedesse todo o treinamento com os demais indivíduos e o primeiro indivíduo apenas sendo testado contruindo-se uma

função discriminante para essas observações. Em seguida deixa-se para teste o segundo indivíduo e treinam-se os restantes e assim sucessivamente até que a melhor Função Discriminante seja ajustada para o total dos casos em ambos os grupos. Dessa forma pode-se calcular a probabilidade de classificações erradas com maior garantia, embora tal procedimento possa 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. São Paulo, 2002.

ANDERSON, T.W. **A introduction to multivariate statistical analysis**. Nova York: Wiley, 1984.

AZEVEDO, LEONARDO VIEIRA. **Maximizando o valor do relacionamento com o cliente: *Data Mining* e CRM**. Disponível em: <http://www.wgsystemens.com.br>Acesso em 12 dez. 2004.

BÄCK, T.; HAMMEL, U.; SCHWEFEL, H. **Evolutionary Computation Comments on the history and Current State**. IEEE Transactions on Evolutionary Computation, v.1, n.1, 1997.

CANUTO, ANNE MAGALY de PAULA; GOTTGTROY, MÁRCIA de PAIVA BASTOS. ***Data Mining*: Geração de dados com qualidade para sistemas agropecuários**. Disponível em: <www.agrosoft.org.br >Acesso em: 05 dez.2004.

CARVALHO, ANTONIO VIEIRA DE ; NASCIMENTO , LUIZ PAULO. **Administração de recursos humanos**.São Paulo, v.1.Pioneira, 1999.

CARVALHO, A. P. de L. F. **Redes Neurais artificiais**. Disponível:em:<http://www.icmc.sc.usp.br >Acesso em: 20 nov.2004.

CHIARA, RAMON. **Aplicação de *Data Mining* em logs de sevidores web. mht.** Instituto de Ciências e de Computação. Dissertações de mestrado da USP. Disponível em: <www.saber.usp.br > Acesso em 1 nov.2004.

CHIAVENATO IDALBERTO - **Gestão de pessoas: O novo papel dos recursos humanos nas organizações**. Rio de Janeiro: Elsevier, 1999.

DILLY, R. **Data Mining – An Introduction**. Belfast: Parallel Computer Center, Queens

University, 1999.

DRAPPER, N. R.; SMITH, H. ***Apllied Regression Analysis***. 2ª Ed. Nova York: Wiley, 1981.

FAYYAD,U. M; PIATETSKY-SHAPIRO,G; SMYTH, P; THURUSAMY,R. ***Advances in Knowledge Discovery & Data Mining.*** Cambridge, MA: AAAI/MIT, 1996.

FISHER, R.A. The statistical utilization of multiple measurements. Annals of Eugenics, 1938.

GOLDEBERG, DAVI E. Genetic algoritmics in search, optimzation, and machine learning. New York, 1989.

GOMES, JOSIR CARDOSO; LEVY, ARIEL; LACHTERMACHER GERSON. Segmentação do censo educacional 2000 utilizando técnicas de mineração de dados. Simpósio Brasileiro de Pesquisa Operacional. São João Del Rey. **Anais...**São João Del Rey, 2004.p 57.

GUIDE. Guide to *Data Mining*. **Introdução ao *Data Mining***. Disponível em: < http://www.data-mining-guide.net/> Acesso em: 07 mar. 2005.

GUIMARÃES, ALAINE MARGARETE. Inteligência computacional aplicada a *Data Mining*. ERI 2004, XII Escola Regional de Informática. Guarapuava. **Anais...**Guarapuava, 2004. p.90-132.

HAYKIN, SIMON. **Redes Neurais: Princípios e práticas**. 2ª. Porto Alegre: Bookman, 2001

HERRERA, F. ***Tackling real-coded genetic algorithms***: **operators and tools for behavioral *analysis***, 1996.

HOLSHEMIER M.; SIEBES A. ***Data Mining***. Disponível em: <http://www.goldnet.it/~daniele/node19.html > Acesso em 07 mar. 2005.

HOSMER, D.W.; LEMESHOW, S. ***Applied Logistic Regression.*** Nova York: Wiley,1989.

INMON, W.H. ***Building the Data Warehouse***, Nova York: Wiley, 1993.

INMON, W.H. **The Data Warehouse and Data Mining**. Communication of the ACM, 1996. p.40-50.

JOHNSON, R.A. e WICHERN, D.W. ***Applieed multivariate statistical analysis***. Londres:Prentice-Hall, 1982.

KLEINBAUM D.G; KUPPER L.; MULLER K; NIZAM A. ***Applied regression analysis and other multivariable methods***. 3.ed. Boston: Duxbury Press, 1998.

KÓVACS, ZSOLT L. **Redes neurais artificiais: fundamentos e aplicações**. 3.ed. São Paulo: Livraria da Física, 2002.

LOUZADA NETO, FRANCISCO; DINIZ, CARLOS ALBERTO RIBEIRO. **Técnicas estatísticas em *Data Mining***. Monografias Del IMCA nº 31. Lima-Peru: IMCA,2002.

MARQUES, JAIR MENDES. Notas de aula da disciplina de Análise Multivariada Aplicada a Pesquisa, do curso de Mestrado Interestitucional em Métodos Numéricos em Engenharia. Guarapuava, Fev.2004.

NAVEGA, SÉRGIO. Princípios essenciais do *Data Mining*. Infoimagem, Cenadem 2002. São Paulo.**Anais**...São Paulo, 2002. Disponível em: <http://www.intelliwise.com/snavega>. Acesso em: 20 jul. 2005.

NIEVOLA, JÚLIO CESAR. Redes neurais artificiais. ERI 2004, XII Escola Regional de Informática. Guarapuava. **Anais**...Guarapuava, 2004. p.01-50.

NIMER, F; SPANDRI, L.C. *Data Mining*. **Revista *Developers***. Fev.1998, p32.

NOTARI, DANIEL LUÍS. **Aplicação de redes neurais artificiais à mineração de dados**: Monografia de Conclusão para a Obtenção do Grau de Bacharel em Ciência da Computação na UCS. Caxias do Sul, Dez. 1997.

PALAZZO, L. **Algoritmos para computação evolutiva**. Universidade Católica De Pelotas –

Escola de Informática.Grupo de Pesquisa em Inteligência Artificial. Disponível em: <www.ucpel.tche.br.> Acesso em: 05 mar. 2004.

PAULA DE, G.G. JÚNIOR; SILVEIRA, M. R. M.; FONSECA, R. NETO.Uma rede neural artificial de múltiplas camadas aplicada ao combate à sonegação fiscal de icms. Simpósio Brasileiro de Pesquisa Operacional. São João Del Rey. **Anais**...São João Del Rey, 2004.p 78.

PARPINELLI, R. S. e LOPES, H.S. E FREITAS, A. A. **Data Mining with a Ant Colony Optimization Algorithm**. IEEE Trans. on Evolutionary Computation, special issue on Ant Colony Algorithms, 2002.

PETER, CHEESMAN; ROBIN, HANSON; JOHN, STUTZ. **Bayesian classification with correlation and inheritance**: In 12th International Conference on Artificial Intelligence. Ago. 1991.

QUEIROZ, A. E. de M.; GOMES, A.;CARVALHO, F. de A. T. **Mineração de dados de IHC para interface educativas**. UFPE - Centro de Informática. Disponível em: <http://www.sbc.org.br/reic/edicoes/2002e4/cientificos/MineracaoDeDadosDeIHCParaInterfacesEducativas.pdf > Acesso em 23/04/2004.

RAMOS, B. P. PEREIRA; LEONCINI, F. L.; YWAMOTO, G. K.;LIZIÉR, M. A. S.; COLOMBINI, T. M.. **Descoberta de conhecimento em base de Dados**. Monografia de Bacharelado em Ciências da Computação do IMC, USP, São Carlos, 2003.

RODRIGUES, ALEXANDRE MEDEIROS, **Técnicas de *Data Mining* classificadas do ponto de vista do usuário**. Dissertação de mestrado em engenharia de produção da UFRJ. Rio de Janeiro, 2000.

SILVA FILHO, ABRANTES ARAÚJO. **Resumo sobre regressão logística**. ENSP/FIO CRUZ. Disponível em : < www.ensp.fiocruz.br/ > Acesso em 05 jul. 2005.

SILVA, MABEL PEREIRA; BOSCARIOLI, CLÓDIS E PERES SARAJANE MARQUES. **Análise de logs da web por meio de técnicas de *Data Mining***. Disponível em: <www.unioeste.com > Acesso em 19 ago. 2004.

SOARES, PEDRO PAULO SILVA. Aplicação de uma rede neural feedforward com algoritmo Levenberg-Marquardt para classificação de laterações do segmento ST do eletrocardiograma. IV Congresso Brasileiro de Redes Neurais, 1999, São José dos Campos. **Anais**...São José dos Campos, 1999. p 384-389.

STEINER, M.T.A. Notas de aula da disciplina de Programação Inteira e Otimização em Redes, do curso de Mestrado Interestitucional em Métodos Numéricos em Engenharia. Guarapuava, jun. 2004.

VIANA REINALDO. Mineração de dados: introdução e aplicações.**Revista SQL Magazine**, Rio de Janeiro: Neofício, n.10/1, 2004.

APÊNDICE 1 -QUESTIONÁRIO

CÓDIGO DE IDENTIFICAÇÃO: LOJA : CÓD.:

01) Idade?	
1	Entre 18 e 22 anos
2	Entre 22 e 28 anos
3	Entre 28 e 34 anos
4	Entre 34 e 40 anos
5	Entre 40 e 46 anos
6	Entre 46 e 52 anos
7	Entre 52 e 60 anos
8	Acima de 60 anos.
02)Sexo	
1	masculino
2	feminino
03)Estado Civil	
1	solteiro
2	casado
3	viúvo
4	desquitado
5	outro
04)Número de filhos	
0	nenhum filho
1	1 filho
2	2 filhos
3	3 filhos
4	4 filhos
5	5 filhos
6	6 filhos
7	mais que 6 filhos
05)Em sua família você é	
1	primeiro filho
2	segundo filho
3	terceiro filho
4	quarto filho
5	quinto filho
6	sexto filho
7	acima do sexto
06) Grau de instrução	
1	até 4a. Série
2	até 6a. Série
3	até 8a. Série
4	ensino médio incompleto
5	ensino médio completo
6	superior incompleto
7	superior completo

07)Religião	
1	Católico
2	Evangélico
3	Protestante
4	Adventista
5	Judia
6	Muçulmana
7	Espírita
8	Budista
9	Outra .....
08)Qual o mês do seu nascimento	
1	Janeiro
2	Fevereiro
3	Março
4	Abril
5	Maio
6	Junho
7	Julho
8	Agosto
9	Setembro
10	Outubro
11	Novembro
12	Dezembro
09) Animal preferido	
1	Cachorro
2	Gato
3	Pássaro
4	Cavalo
5	Vaca
6	Carneiro
7	Porco
8	Leão
9	Outro.....
10) Cor preferida	
1	Amarelo
2	Azul
3	Verde
4	Vermelho
5	Preto
6	Branco
7	Alaranjado
8	Violeta
9	Cinza

11)Seu tempo de folga é ocupado com:	
1	leitura ou estudo
2	tarefas domésticas
3	esporte
4	encontro com amigos
5	esporte
6	internet
7	dedicação a família
8	viajar
9	dormir
10	outro.....
12) Indique no espaço o número do setor em que você trabalha verificando a lista de setores com o gerente da loja	
13) Indique no espaço o número da sua função verificando a lista com o gerente da loja	
14) Há quanto tempo você faz parte do quadro de colaboradores da empresa?	
1	em experiência
2	menos de 3 meses
3	entre 3 meses e 6 meses
4	entre 6 meses e 1 ano
5	entre 1 ano e 2 anos
6	entre 2 anos e 3 anos
7	entre 3 anos e 4 anos
8	entre 4 anos e 5 anos
9	entre 5 anos e 7 anos
10	entre 7 anos e 10 anos
11	acima de 10 anos
15) Qual a distância de sua residência até o trabalho?	
1	até 500 metros
2	de 501 à 1000 metros
3	de 1001 à 2000 metros
4	de 2001 à 3000 metros
5	de 3001 à 4000 metros
6	de 4001 à 5000 metros
7	de 5001 à 8000 metros
8	acima de 8001 metros

16) Você vai ao trabalho:

1	à pé
2	de bicicleta
3	de moto
4	de ônibus
5	de carro próprio
6	de carona
7	com veículo da empresa

17)Quantos trabalhos com registro em carteira você já teve?

1	Único(Atual)
2	Dois
3	Três
4	Quatro
5	Cinco
6	Seis
7	Mais de seis

18)Quanto a sua função você está..

1	totalmente satisfeito
2	parcialmente satisfeito
3	insatisfeito
4	gostaria de estar em outra
	função

19)Quanto a loja que voce trabalha, você está...

1	totalmente satisfeito
2	parcialmente satisfeito
3	insatisfeito
4	gostaria de estar em outra
	loja

20) Seu relacionamento com seus colegas.

1	ótimo
2	bom
3	regular
4	ruim
5	péssimo
6	não possui

21) Seu relacionamento com seus superiores diretos é:

1	ótimo
2	bom
3	regular
4	ruim
5	péssimo
6	não possui

22) Seu relacionamento com a direção da empresa é:

1	ótimo
2	bom
3	regular
4	ruim
5	péssimo
6	não possui

23) Indique sua faixa salarial:

1	entre R$300,00 e R$500,00
2	entre R$501,00 e R$700,00
3	entre R$701,00 e R$900,00
4	entre R$901,00 e R$1.200,00
5	entre R$1.201,00 e R$1.500,00
6	entre R$1.501,00 e R$2.000,00
7	entre R$2.001,00 e R$2.500,00
8	acima de R$2.501,00

24)Diante da situação atual do mercado de trabalho em sua cidade e em comparação a outras pessoas, conhecidas suas, você julga seu nível salarial

1	ótimo
2	bom
3	regular
4	ruim
5	péssimo

25)Se você pudesse trabalhar em outra empresa de sua cidade, com o mesmo nível salarial que você possui hoje, você

1	continuaria em seu trabalho
2	mudaria para o outro ofertado

26)Quanto tempo de trabalho você acha ideal para um colaborador trabalhar na mesma empresa?

1	menos de 1 ano
2	entre 1 e 2 anos
3	entre 2 e 3 anos
4	entre 3 e 5 anos
5	entre 5 e 8 anos
6	entre 8 e 12 anos
7	acima de 12 anos

27) O principal motivo que levaria você solicitar sua rescisão de contrato de trabalho

1	desentendimento c/ colegas
2	desentendimento c/ superiores
3	melhor proposta salarial
4	continuidade dos estudos
5	problemas familiares
6	liberação de fgts
7	stress e problemas de saude
8	não pediria rescisão

28) De uma maneira geral você considera seu trabalho:

1	ótimo
2	bom
3	regular
4	ruim
5	péssimo

29) De maneira geral você considera a empresa :

1	ótima
2	boa
3	regular
4	ruim
5	péssima

30) Como você prevê a oferta de emprego em nosso país nos próximos 5 anos

1	Acredito que vai melhorar
2	Não acredito em melhorias

APÊNDICE 2 – AUTOVALORES

Autovalores da matriz de Correlação			
		variância	var.explicada
variável	autovalor	explicada %	acumulada %
1	2,6492	12,62	12,62
2	2,3450	11,17	23,78
3	1,6084	7,66	31,44
4	1,2891	6,14	37,58
5	1,1671	5,56	43,14
6	1,1132	5,30	48,44
7	1,0418	4,96	53,40
8	1,0268	4,89	58,29
9	0,9965	4,75	63,03
10	0,9380	4,47	67,50
11	0,9092	4,33	71,83
12	0,8786	4,18	76,01
13	0,8461	4,03	80,04
14	0,8395	4,00	84,04
15	0,6783	3,23	87,27
16	0,6400	3,05	90,32
17	0,5950	2,83	93,15
18	0,5642	2,69	95,84
19	0,4248	2,02	97,86
20	0,3432	1,63	99,49
21	0,1063	0,51	100,00

APÊNDICE 3 - COMUNALIDADES

variáveis	até o 10 fator	até o 11 fator	até o 12 fator	até o 13 fator	até o 14 fator	R multiplo
VAR1	0,67371154	0,71712018	0,7246859	0,75207428	**0,75342374**	0,51207903
VAR2	0,7123093	0,71256693	0,72645331	0,74731711	**0,78381302**	0,19906045
VAR3	0,61919992	0,61928431	0,61928456	0,62439105	**0,6247961**	0,22530503
VAR4	0,97718957	0,97779696	0,97791961	0,97809425	**0,97831891**	0,03970031
VAR5	0,01084115	0,01174773	0,01210031	0,95808114	**0,95954271**	0,06173185
VAR6	0,18352219	0,212745	0,21746674	0,23380312	**0,81130369**	0,21649393
VAR7	0,97634175	0,97643166	0,97644864	0,97827778	**0,97830488**	0,03679814
VAR8	0,98949483	0,98957731	0,98983943	0,99024324	**0,99039358**	0,0139825
VAR9	0,94499015	0,94515829	0,94585309	0,94625796	**0,94767017**	0,05974613
VAR10	0,00647886	0,00648283	0,97735091	0,97769535	**0,97788594**	0,03601056
VAR11	0,97719669	0,97921522	0,97922816	0,98102959	**0,98154929**	0,03979693
VAR12	0,56190918	0,5648619	0,56690939	0,58311947	**0,64997447**	0,20646597
VAR13	0,01517563	0,91183637	0,91186129	0,91276171	**0,91374609**	0,09424756
VAR14	0,8064597	0,81489849	0,82019442	0,82021341	**0,83050965**	0,27453488
VAR15	0,70251585	0,72952908	0,73682946	0,74052172	**0,77526851**	0,37979588
VAR16	0,6410998	0,64826449	0,64839158	0,64885841	**0,66039274**	0,34170026
VAR17	0,31607237	0,38697048	0,39222324	0,39874582	**0,75349458**	0,30679408
VAR18	0,92916582	0,92955697	0,93205719	0,93214371	**0,93495217**	0,05513387
VAR19	0,6370945	0,64027595	0,64548756	0,64560371	**0,65210024**	0,57129804
VAR20	0,78344951	0,78936903	0,79041151	0,79043541	**0,79103281**	0,72760387
VAR21	0,92538503	0,92877835	0,92902862	0,9291393	**0,9294406**	0,81776008

APÊNDICE 4 - RESÍDUOS

Correlações Residuais

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10	VAR11
VAR1	0,25398	0,00569	-0,15163	-0,00057	-0,04063	0,02905	0,02403	0,00933	0,01347	-0,0231	0,00786
VAR2	0,00569	0,22445	-0,11052	0,00757	-0,06555	-0,13072	0,01276	0,03209	0,01001	-0,07449	0,03652
VAR3	-0,15163	-0,11052	0,35982	0,02291	0,05235	-0,00706	-0,06334	-0,01779	-0,01298	0,05053	-0,02414
VAR4	-0,00057	0,00757	0,02291	0,02113	-0,00065	0,00472	-0,00557	0,00089	-0,01832	0,01302	-0,00213
VAR5	-0,04063	-0,06555	0,05235	-0,00065	0,02855	0,05223	-0,00684	-0,01116	-0,0025	0,02891	-0,01319
VAR6	0,02905	-0,13072	-0,00706	0,00472	0,05223	0,19711	0,00844	-0,02195	-0,00199	0,06605	-0,02844
VAR7	0,02403	0,01276	-0,06334	-0,00557	-0,00684	0,00844	0,01431	0,00093	-0,00133	-0,00725	0,00355
VAR8	0,00933	0,03209	-0,01779	0,00089	-0,01116	-0,02195	0,00093	0,00543	0,00358	-0,01205	0,00554
VAR9	0,01347	0,01001	-0,01298	-0,01832	-0,0025	-0,00199	-0,00133	0,00358	0,03262	-0,01718	0,00323
VAR10	-0,0231	-0,07449	0,05053	0,01302	0,02891	0,06605	-0,00725	-0,01205	-0,01718	0,04152	-0,01627
VAR11	0,00786	0,03652	-0,02414	-0,00213	-0,01319	-0,02844	0,00355	0,00554	0,00323	-0,01627	0,00716
VAR12	-0,01789	0,17761	0,05286	0,04629	-0,03816	-0,05215	-0,00449	0,01909	-0,02674	-0,02841	0,02074
VAR13	-0,05353	-0,00283	0,04876	0,01159	0,02407	0,07316	-0,00927	-0,00363	0,00759	0,02432	-0,00905
VAR14	0,0079	-0,02911	0,00658	0,04991	0,0156	0,0697	-0,01316	-0,00162	-0,03505	0,05624	-0,01578
VAR15	-0,00506	-0,0144	0,00277	-0,04134	-0,00606	-0,04424	0,0197	-0,00738	0,00589	-0,03569	0,00913
VAR16	-0,07381	0,00262	-0,20478	-0,02292	0,02629	0,12247	0,04806	-0,00952	-0,00267	0,0177	-0,00405
VAR17	-0,03193	-0,00319	-0,04169	-0,0094	-0,01657	-0,12428	-0,00619	0,00635	-0,00927	-0,01448	0,00633
VAR18	-0,00332	-0,05472	-0,01265	0,01236	0,02215	0,08042	0,00825	-0,01111	-0,0239	0,03683	-0,01269
VAR19	0,01197	-0,03147	0,00998	0,04598	-0,00508	-0,01139	0,01297	-0,01036	-0,09935	0,0374	-0,00511
VAR20	0,00402	-0,04085	-0,02753	0,03077	0,0011	0,00397	0,01644	-0,00976	-0,07367	0,03312	-0,006
VAR21	0,00908	0,01896	0,01821	0,01069	-0,00522	-0,00843	-0,00773	0,00498	0,00577	-0,00176	0,0013

	VAR12	VAR13	VAR14	VAR15	VAR16	VAR17	VAR18	VAR19	VAR20	VAR21
VAR2	-0,01789	-0,05353	0,0079	-0,00506	-0,07381	-0,03193	-0,00332	0,01197	0,00402	0,00908
VAR3	0,17761	-0,00283	-0,02911	-0,0144	0,00262	-0,00319	-0,05472	-0,03147	-0,04085	0,01896
VAR4	0,05286	0,04876	0,00658	0,00277	-0,20478	-0,04169	-0,01265	0,00998	-0,02753	0,01821
VAR5	0,04629	0,01159	0,04991	-0,04134	-0,02292	-0,0094	0,01236	0,04598	0,03077	0,01069
VAR6	-0,03816	0,02407	0,0156	-0,00606	0,02629	-0,01657	0,02215	-0,00508	0,0011	-0,00522
VAR7	-0,05215	0,07316	0,0697	-0,04424	0,12247	-0,12428	0,08042	-0,01139	0,00397	-0,00843
VAR8	-0,00449	-0,00927	-0,01316	0,0197	0,04806	-0,00619	0,00825	0,01297	0,01644	-0,00773
VAR9	0,01909	-0,00363	-0,00162	-0,00738	-0,00952	0,00635	-0,01111	-0,01036	-0,00976	0,00498
VAR10	-0,02674	0,00759	-0,03505	0,00589	-0,00267	-0,00927	-0,0239	-0,09935	-0,07367	0,00577
VAR11	-0,02841	0,02432	0,05624	-0,03569	0,0177	-0,01448	0,03683	0,0374	0,03312	-0,00176
VAR12	0,02074	-0,00905	-0,01578	0,00913	-0,00405	0,00633	-0,01269	-0,00511	-0,006	0,0013
VAR13	0,34508	0,06319	0,00442	-0,00503	-0,06218	-0,17821	-0,0094	0,071	0,00591	0,03063
VAR14	0,06319	0,0816	0,04323	-0,05717	0,05233	-0,09718	0,02265	-0,05954	-0,04825	0,01496
VAR15	0,00442	0,04323	0,19706	-0,19032	0,0074	0,02911	0,04985	0,06384	0,05394	0,02309
VAR16	-0,00503	-0,05717	-0,19032	0,22297	0,00752	-0,03894	-0,01849	0,03478	0,02564	-0,03431
VAR17	-0,06218	0,05233	0,0074	0,00752	0,36052	-0,03931	0,07487	-0,01897	0,03435	-0,03567
VAR18	-0,17821	-0,09718	0,02911	-0,03894	-0,03931	0,26697	-0,03696	0,03185	0,04762	-0,01186
VAR19	-0,0094	0,02265	0,04985	-0,01849	0,07487	-0,03696	0,05163	0,06315	0,06099	-0,00582
VAR20	0,071	-0,05954	0,06384	0,03478	-0,01897	0,03185	0,06315	0,34093	0,22244	-0,05164
VAR21	0,00591	-0,04825	0,05394	0,02564	0,03435	0,04762	0,06099	0,22244	0,22567	0,0194

ANEXO 1 - SETORES

Código	Setores
1	Administração
2	Açougue
3	Padaria
4	Rotisseria
5	Depósito
6	Foto/Vídeo
7	Hortifruti
8	Frente de Caixa
9	F.L.C.
10	Padaria/Confeitaria/Rotisseria
11	Reposição
12	Lotérica
13	Transportes
14	Restaurante
15	Confeitaria
16	Contabilidade
17	Financeiro
18	Recursos Humanos
19	Assistente Gerencial
20	Controladoria
21	Diretoria
22	Conservação e Limpeza
23	Bazar Pesado
24	Informática
25	Compras
26	C.P.D.
27	Higiênico Sanitário
28	Segurança
29	Bazar Leve
30	Peixaria
31	Central Produção
32	Central de Trocas
33	Faturamento
34	Entrada de Mercadorias
35	Tesouraria
36	Estacionamento
37	Marketing
38	Controle de Estoques
39	Guarda-volumes
40	Pesquisa
41	Prevenção Perdas
42	Salgadaria
43	Frios/Padaria/Restaurante

ANEXO 2 - FUNÇÕES

cód	Função	cód.	Função	Cód.	Função
1	Açougueiro	98	Enc. Admin. de Pessoal	151	Locutor
109	Assist.Auditoria	24	Enc. Compras	148	Médico do Trabalho
78	Assist. Compras	25	Enc. Confeitaria	67	Motorista Kombi
2	Assist. Contábil	42	Enc. Cozinha	69	Motorista Truck
5	Assist. Dpto. Pessoal	26	Enc. Depósito	52	Operador de Caixa
96	Assist. Financeiro	28	Enc. Feira	152	Operador de Loja
3	Assist. Gerencial	104	Enc. Foto/Vídeo	128	Operador de Empilhadeira
166	Assist. Hig. Sanitario	30	Enc. Frente de Caixa	53	Padeiro
103	Assist. Informática	32	Enc. Frios/Lact.	80	Receb. Mercadorias
4	Assist. Loja	33	Enc. Limpeza	56	Recepcionista
6	Assist. Preços	35	Enc. Peixaria	57	Repositor
172	Assist.de depósitos	36	Enc. Prod. De Pães	76	Repositor de Frios
102	Atendente	37	Enc. Reposição	70	Salgadeiro
8	Aux. Açougue	91	Enc. Setor Contábil	149	Técnico em Enfermagem
111	Aux. Administrativo	101	Enc. Setor de Controladoria	156	Tesoureiro
114	Aux. Alimentação I	29	Enc. Setor Financeiro	72	Vendedor Atacado
115	Aux. Alimentação II	31	Enc.Trein. Rec. Seleção	73	Zelador
116	Aux. Cadastro	41	Enc. Vendas Pães		
121	Aux. Caixa	112	Encarregado		
77	Aux. Comercial	147	Engenheiro de Segurança		
10	Aux. Confeitaria	43	Escriturário		
92	Aux. Contábil	44	Estagiário		
11	Aux. Cozinha	154	Faturista		
12	Aux. Depósito	45	Feirante		
16	Aux. Depto. Pessoal	79	Fiscal de Caixa		
126	Aux. Entrega	46	Fiscal de Loja		
87	Aux. Estoque	123	Fiscal de Portaria		
170	Aux. Faturamento	129	Fiscal de Portaria		
81	Aux. Financeiro	100	Gerente Adm/Financeiro		
93	Aux. Fiscal	48	Gerente Administrativo		
14	Aux. Frente de Caixa	158	Gerente Adm. Trainee		
145	Aux. Informática	99	Gerente Comercial		
113	Aux. Limpeza	171	Gerente de Bazar		
117	Aux. Loja I	168	Gerente de Informática		
118	Aux. Loja II	161	Gerente de Mercearia		
90	Aux. Motorista	157	Gerente de Merc.-Trainee		
15	Aux. Padaria	167	Gerente de Operações		
75	Aux. Peixaria	159	Gerente de Perecíveis		
18	Balconista	162	Gerente de Prev.e Perdas		
19	Cartazista	163	Ger. de Prev.e Perdas Trainee		
20	Comprador	59	Gerente Geral		
21	Confeiteiro	122	Gerente Geral Trainee		
127	Conferente	60	Gerente operacional		
64	Control. Estacion.	160	Gerente Perecíveis Trainee		
22	Controle Estoque	61	Gerente de RH.		
74	Cozinheiro	63	Gerente de Vendas		
23	Enc. Açougue	65	Guarda Volumes		

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo