PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS

CARLA REGINA MAZIA ROSA

UMA METODOLOGIA PARA A DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS VISANDO A CLASSIFICAÇÃO DE PADRÕES

CURITIBA
2017

CARLA REGINA MAZIA ROSA

UMA METODOLOGIA PARA A DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS VISANDO A CLASSIFICAÇÃO DE PADRÕES

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de doutor em Engenharia de Produção.

Orientadora: Prof.ª Dr.ª Maria Teresinha Arns Steiner

CURITIBA

2017

Rosa, Carla Regina Mazia

R788m 2017

Uma metodologia para a descoberta de conhecimento em bases de dados visando a classificação de padrões / Carla Regina Mazia Rosa ; orientadora, Maria Teresinha Arns Steiner. – 2017.

161 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2017.

Bibliografia: f. 147-158

1. Mineração de dados (Computação. 2. Redes neurais. (Computação). 3. Análise multivariada. 4. Tecnologia médica. I. Steiner, Maria Teresinha Arns. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Engenharia de Produção e Sistemas. III. Título.

CDD 20. ed. – 006.312

CARLA REGINA MAZIA ROSA

UMA METODOLOGIA PARA A DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS VISANDO A CLASSIFICAÇÃO DE PADRÕES

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de doutor em Engenharia de Produção.

COMISSÃO EXAMINADORA

Professora Dr.ª Maria Teresinha Arns Steiner
Pontifícia Universidade Católica do Paraná

Professor Dr. Pedro José Steiner Neto
Universidade Positivo

Professor Dr. Júlio Nievola
Pontifícia Universidade Católica do Paraná

Professor Dr. Anderson Roges Teixeira Góes
Universidade Federal do Paraná

Professor Dr. Deise Maria Bertholdi Costa
Universidade Federal do Paraná

Curitiba, 22 de fevereiro de 2017

AGRADECIMENTOS

- Agradeço primeiramente, à Deus por ter me dado a graça de ter a sabedoria de aprender a cada obstáculo. Cada um deles ao seu modo, me fizeram chegar onde eu cheguei;
- Agradeço eternamente, ao meu marido Robson Souza Rosa, pelo amor e, principalmente, por acumular muitas das minhas atividades domésticas e por incentivar-me em todos os momentos, em especial, naqueles em que a força e a determinação já estavam se esgotando;
- Agradeço, em especial ao meu filho amado Henry Mazia Rosa, pelo amor incondicional, carinho e espontaneidade que sempre me estimularam nos momentos difíceis, mesmo que ainda não tem idade para entender o que é uma tese e quão difícil é;
- Agradeço, à minha família que me incentivou nessa jornada e por compreender os momentos de minha ausência;
- Agradeço à minha orientadora, professora Dr.ª Maria Teresinha Arns Steiner, pela confiança depositada, por todo o inestimável conhecimento transmitido;
- Agradeço, à minha amiga querida Cleina Yayoe Okoshi, pelas longas conversas e por sempre estar disposta a me ajudar;
- Agradeço, ao Professor Dr. Pedro José Steiner Neto, pela contribuição dada para o meu aprendizado na análise dos dados;
- Agradeço, ao Hospital das Clínicas de Curitiba, em especial ao Dr. Clóvis Roehrig, pelo fornecimento dos dados para a realização deste trabalho;
- Agradeço à instituição de ensino superior privada, que permitiu a realização da coleta de dados em seu estabelecimento;
- Agradeço à empresa que permitiu o acesso ao seu processo;
- Agradeço aos meus amigos e colegas do grupo de pesquisa e a todos aqueles que, de alguma maneira, contribuíram para a realização deste trabalho;
- Agradeço à Secretaria do Curso, pela cooperação;
- Agradeço à CAPES, pelo apoio financeiro.

RESUMO

O presente trabalho tem por objetivo apresentar uma análise exploratória dos dados e seus impactos nas técnicas de Mineração de Dados para a classificação de padrões (instâncias). A referida metodologia detecta quais são os atributos mais importantes e realiza a classificação de novas instâncias de forma automatizada e com a máxima acurácia. A mesma está sendo aplicada a três problemas reais, com o intuito de testar as suas diversas etapas: nos cursos de Pós-Graduação Lato Sensu de uma Instituição de Ensino Superior privada, no diagnóstico médico e, finalmente, na empresa de calibração de balança. Com base no processo Descoberta de Conhecimento em Bases de Dados foi realizada inicialmente, uma análise exploratória dos dados, com o intuito de analisar se há diferença entre os atributos de cada classe e, além disso, detectar atributos atípicos e, em seguida, as técnicas de *Data Mining*, com a finalidade de avaliar os seus desempenhos quanto a classificação de novas instâncias. As técnicas de *Data Mining* utilizadas foram: Regressão Logística Binária; Geração de uma Superfície que Minimiza Erros, Função Discriminante Linear de Fisher, Redes Neurais Artificiais e Máquina de Vetores de Suporte, sendo que a primeira e terceira são técnicas estatísticas e a segunda, quarta e quinta fazem uso da Programação Linear. Por meio dos resultados obtidos, pode-se observar que a técnica de Redes Neurais Artificiais, precedida de uma análise exploratória, foi a que apresentou os melhores resultados para o primeiro problema, com uma taxa de acerto de 92,54%. Da mesma forma, para o problema médico, a Redes Neurais Artificiais também foi a que apresentou a melhor acurácia, com uma taxa de 90,68%. Já para o problema de calibração de balança, a Função Discriminante Linear de Fisher foi a que apresentou o melhor desempenho, com uma taxa de acerto de 60,39%. Assim, dadas novas instâncias, o especialista teria um respaldo adicional para realizar o seu diagnóstico para os problemas abordados, com base nas técnicas “já treinadas”.

Palavras-chave: Descoberta de Conhecimento em base de Dados. Mineração de Dados. Análise Exploratória dos Dados. Reconhecimento de Padrões.

ABSTRACT

This paper aims at introducing a method for discovering knowledge in databases that can be used in several different areas for pattern recognition (instances). This method detects which variables (attributes; information) are the most important ones and performs the ranking of new instances in an automatized way with maximum accuracy. This method is being applied to three real problems aiming at testing its several stages: to a problem in Graduate Education Lato Sensu courses at a private Higher Education Institution in which it is intended to check the satisfaction and quality of services provided, in a medical problem aiming at classifying cholestatic patients and, finally, in a problem of calibrating weighing scale with the intention of analyzing the quality of services provided. For the first problem (Lato Sensu Graduate courses), 885 pieces of data were collected, each with 12 variables and one exit (satisfaction). For the second problem (medical diagnosis), 118 pieces of data were collected, each with 14 variables and one exit (cancer or gallstones). As for the third problem (weighing scale calibration), 1540 pieces of data were collected, each with 14 variables and one exit (good or low quality). From Knowledge Discovery in Databases, an exploratory analysis of data was carried out. It was intended on analyzing whether or not there is a difference between each class attributes and, besides that, on detecting unusual attributes. After that, Data Mining techniques were used to assess performance of KDD (Knowledge Discovery in Databases) in regards to classification of new instances. The following Data Mining techniques were used: Binary Logistic Regression, Generation of a Surface that Minimizes Errors, Fisher Linear Discriminant, and Support Vector Artificial Neural Networks; the first and the third are statistics techniques and the second, fourth, and fifth ones use Linear Programing. Considering the obtained results, what can be seen is that Artificial Neural Networks preceded by an exploratory analysis was the one that presented the best results for the first problem with a success rate of 92.54%. In a similar fashion, for the medical problem, Artificial Neural Networks was also the one that presented the best accuracy with a rate of 90.68%. As for the weighing scale problem, Fisher Linear Discriminating Function was the one that presented best performance with a success rate of 60.39%. Therefore, given the new instances, the expert would have additional background to carry out their diagnosis for the approached problems based on the techniques “already trained”.

Key-words: Knowledge Discovery in Data Bases. Data Mining. Binary Logistic Regression. Generating a surface that minimizes errors. Fisher Linear Discriminant function. Artificial Neural Networks. Support Vector Machine.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACP	Análise de Componentes Principais
AG	Algoritmo Genético
AGs	Algoritmos Genéticos
ABRF	Avaliação Bioquímica de Risco Fetal
BAS	Building Automation System / Sistema de Automação Predial
CART	*Classification and Regression Trees*
CRISP-DM	*CRoss Industry Standard Process for Data Mining* / Processo Padrão Inter Indústrias para Mineração de Dados
DCPG	*Distributed Computation Precedence Graph* / Computação Distribuída de Precedência Gráfica
DM	*Data Mining*/Mineração de dados
EFSS	*Evolutionary Fuzzy Systems* / Sistemas Nebulosos Evolutivos
FDLF	Função Discriminante Linear de Fisher
FFD	Detecção de Fraudes Financeiras
GAMS	*General Algebraic Modeling Systems* / Sistema Geral de *Modelagem Algébrica*
GLIM	*Generalized Linear Interactive Model* / M*odelo Linear Generalizado Interativo*
GSLP-PL	Geração de uma Superfície Linear por Partes
GSME-PL	Geração de uma Superfície que Minimiza Erros
GPA	*Grade Point Averages* / Média de Notas
GPUS	*Graphics Processing Units* / Unidades de Processamento Gráfico
HC	Hospital das Clínicas
IB	*Instance-Based* / Baseada em Instância
IDTUV2	*Induction of Decision Tree with restoring Unknown Values* / Indução de Árvores de Decisão com a restauração de valores desconhecidos
IES	Instituição de Ensino Superior
JSM	*John Stuart Mill*
KDD	*Knowledge Discovery in Data Bases* / Descoberta de Conhecimento em base de Dados
KDDVM	*Knowledge Discovery in Databases Virtual Mart*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*

KMNI	*Imputation with K-Nearest Neighbor* / Imputação com K Vizinhos Mais Próximos
K-NN	*K-Nearest Neighbor* / K-Vizinhos mais Próximos
K-NNC	*K-Nearest Neighbour Classifier* / Classificador K-vizinhos mais próximos
MEC	Ministério da Educação Conselho Nacional de Educação
MEC	Ministério da Educação e Cultura
MUSKUP	*Mugla University Student Knowledge Discovery Unit Program*
MVs	*Missing Values* / Valores Faltantes
NPC	*Nearest Prototype Classifier* / Classificador Vizinho Mais Próximo
PEL-C	*Prototype Exemplar Learning Classifier* / Protótipo Aprendizagem Exemplar Classificador
PEP	*Product Emergence Process* / Processo de Surgimento do Produto
PL	Programação Linear
PLM	*Product Lifecycle Management* / Gerenciamento de Ciclo de Vida de Produto
PLS	*Pregnant Leach Solution*
PRMCPL	*Parallel Regularized Multiple-Criteria Linear Programming*
PG	*Pós-graduação*
RFM	Recente, Frequência e Valor Monetário
RLB	Regressão Logística Binária
RMCLP	*Regularized Multiple-Criteria Linear Programming*
RNA	Rede Neural Artificial
RNAs	Redes Neurais Artificiais
RP	Reconhecimento de Padrões
SD	*Subgroup Discovery* / Descoberta de Subgrupos
SGOT	*Transaminase Glutâmico-Oxalacética*
SGPT	*Transaminases Glutâmico-Pirúvicas*
SIG	*Geographical Information Systems* / Sistemas de Informação Geográfica
RBF	*Radial-Basis Function* / Gaussiana de Funções Radias
SVM	*Support Vector Machine* / Máquina de Vetores de Suporte

VIM	Vocabulário Internacional de Termos Fundamentais e Gerais de Metrologia
TOM4D	*Timed Observation Modelling For Diagnosis* / Modelagem da Observação Cronometrada para Diagnóstico

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

O mercado globalizado está a cada dia se aprimorando mais, trazendo consigo novas ferramentas e novos métodos, devido à evolução tecnológica aliada à informatização dos mais variados processos organizacionais que tem gerado uma grande quantidade de dados armazenados. Assim sendo, diversas ferramentas e métodos têm sido propostos, com o objetivo de se extrair informações destes dados.

Nesse contexto, o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*; KDD) tem sido utilizado para extrair conhecimento de grandes bases de dados e tem se mostrado efetivo na gestão de informações, buscando identificar as mais relevantes e transformá-las em conhecimento útil à tomada de decisão.

O processo KDD se dá pelo processo de identificação de padrões válidos, novos, visando à melhoria do entendimento de um determinado problema, do qual a Mineração de Dados (*Data Mining*; DM) pode ser vista como uma parte fundamental do processo (ROSA; STEINER; STEINER NETO, 2016; FONSECA; NAMEN, 2016). O DM é o elemento responsável pela extração do conhecimento contido em um banco de dados.

O DM é capaz de revelar, automaticamente, o conhecimento que está implícito em grande quantidade de informações armazenadas nos bancos de dados de uma organização. As técnicas de DM podem fazer uma análise antecipada dos eventos, possibilitando prever tendências e comportamentos futuros, permitindo aos gestores a tomada de decisões baseada em fatos e não em suposições (CARDOSO; MACHADO, 2008).

A qualidade dos dados é também uma das principais preocupações no processo KDD. A qualidade do conhecimento extraída é estritamente determinada pela qualidade dos dados fornecidos como entrada. O pré-processamento de dados no processo KDD tem a finalidade de melhorar a qualidade dos dados, tendo como objetivo principal a identificação e remoção de problemas presentes nos dados preliminarmente à utilização dos métodos de extração de conhecimento.

A análise exploratória dos dados, que constitui a primeira grande etapa do processo KDD, tem sido considerada crucial, melhorando drasticamente o desempenho das técnicas de DM (ROSA; STEINER; STEINER, NETO, 2016).

Por esse motivo, a qualidade da informação constitui a excelência dos resultados da pesquisa. É indispensável que seja realizada uma análise dos dados para verificar, por exemplo, se as amostras em análise poderão ser discriminadas (testes *t* de *student*; T^2 de Hotelling; ANOVA ou MANOVA, dependendo do tipo de amostra); se há falta de atributos e/ou de instâncias; se há presença de casos atípicos (*outliers*); se as hipóteses associadas às técnicas escolhidas estão sendo adequadamente atendidas; bem como identificar se os eventuais “afastamentos das condições ideais”; se seria interessante agrupar os atributos por meio da Análise Fatorial (AF) ou da Análise dos Componentes Principais (ACP); e assim por diante. A omissão de tais análises poderá comprometer os resultados de todo um trabalho (RIBAS; VIEIRA, 2011).

O presente trabalho tem como foco o desenvolvimento de uma metodologia para a tarefa de classificação de instâncias, cujo objetivo é atribuir um exemplo (registro; instância; padrão) a uma classe, dentre um conjunto de classes pré-definidas, com base nos valores de seus atributos através do Reconhecimento de Padrões (RP) com a máxima acurácia. O desempenho da metodologia proposta foi verificado em três problemas reais: Primeiramente nos cursos de Pós-Graduação (PG) Lato Sensu de uma Instituição de Ensino Superior (IES) privada, com o intuito de medir a satisfação e a qualidade dos serviços prestados, visando a melhoraria de seus pontos mais críticos e, consequentemente, aumentando a retenção de alunos. Desta forma, os coordenadores dos referidos cursos terão condições de saber onde estão “acertando” ou “errando”, além de ter um respaldo adicional para a correta classificação (alunos “satisfeitos” ou “insatisfeitos”) e, consequentemente, a classificação automática de novas instâncias. Segundamente nos exames clínicos de pacientes com câncer ou cálculo no duto biliar, onde o médico terá um respaldo adicional para a correta classificação (“câncer” ou “cálculo” no duto biliar) de seus pacientes. Terceiramente em uma organização que realiza o serviço de calibração de balanças, com o intuito de avaliar a qualidade dos serviços prestados por mês, colaborador e fabricante, bem como a quantidade de discrepâncias por cada um e compará-la com as condições ambientais do local de calibração, além de ter um

respaldo adicional para a correta classificação (calibração de “boa qualidade” ou “baixa qualidade”).

1.2 OBJETIVOS

Os objetivos deste trabalho podem ser discriminados em geral e específicos, conforme a seguir.

1.2.1 Objetivo Geral

O presente trabalho tem por objetivo apresentar uma análise exploratória dos dados e seus impactos nas técnicas de Mineração de Dados. Por meio da referida análise é possível classificar novas instâncias de forma automatizada e com a máxima acurácia.

1.2.2 Objetivos Específico

Para se atingir o objetivo geral, os seguintes objetivos específicos se fizeram necessários:

- Identificar e descrever procedimentos para realizar uma análise exploratória dos dados;
- Identificar e descrever as técnicas de DM em bases de dados;
- Aplicar os procedimentos de análise exploratória dos dados em diferentes áreas de atividade;
- Aplicar e avaliar as técnicas de DM nas bases de dados, entre o “antes” e o “depois” dos procedimentos de análise exploratória dos dados;
- Comparar a performance das técnicas de DM entre o “antes” e o “depois” das aplicações das técnicas de análise exploratórias dos dados.

1.3 JUSTIFICATIVA

As organizações possuem grandes quantidades de informações e cada vez mais necessitam de auxílio técnico computacional apropriado para auxiliar na

análise, na interpretação e no relacionamento dessas informações em busca de conhecimento.

O KDD é uma das áreas de pesquisa altamente dinâmicas, na qual novos métodos e aplicações são propostos a cada dia. O processo KDD inclui ainda análises, interpretações e uso do conhecimento extraído do banco de dados por meio de técnicas de DM.

Um dos maiores desafios do DM é a construção de classificadores precisos e computacionalmente eficientes para bases de dados grandes, em termos de número de instâncias e atributos (HAN; KAMBER; PEI, 2011). Os classificadores são construídos tendo em vista as suas possíveis tarefas, que são, basicamente, realizar a associação; a classificação; a regressão de atributos; o *clustering* (agrupamento) e, finalmente, a visualização de instâncias.

Neste contexto, o presente trabalho concentra-se particularmente na análise exploratória dos dados e na etapa de DM, mais especificamente, nas técnicas de classificação. Existem várias técnicas de classificação que diferem basicamente pela metodologia de tratamento do problema.

Por isto é importante fazer a escolha da técnica a ser usada de forma a buscar aquela que melhor se adapta ao problema. Esta pode ser uma tarefa árdua, exigindo o experimento de várias possibilidades antes de definir qual tipo de modelo que deverá ser usado.

Logo, não basta somente aplicar algumas técnicas de classificação, mas sim, mostrar a necessidade de se analisar os atributos obtidos estatisticamente, para então aplicar as diferentes técnicas e poder verificar qual delas é a mais eficaz para o problema específico.

Assim sendo, é importante descrever informações estruturais dos atributos, cujo processo inclui não apenas a capacidade de designar o padrão a uma classe particular e classificá-lo, mas ao mesmo tempo ter a capacidade de descrever aspectos do atributo para designá-lo a uma outra classe.

1.4 CONTRIBUIÇÕES

O presente trabalho apresenta um diferencial em relação a maior parte dos trabalhos correlatos estudados, pelo fato de apresentar uma análise exploratória dos dados, preliminar à aplicação das técnicas de DM. Tal análise, além de permitir um

melhor entendimento dos dados com os quais se está trabalhando, faz com que haja maior acurácia ao se aplicar as técnicas de DM tornando-se, praticamente, imprescindível.

As técnicas de DM, consideradas tipos mais complexos de função analítica, surgiram com a finalidade de mostrar as informações estratégicas ocultas em bancos de dados, por meio da pesquisa dessas informações e da determinação de padrões e classificações.

Desta forma, a principal contribuição deste trabalho é a proposta de uma metodologia para a tarefa de classificação de instâncias, cujo objetivo é atribuí-las a uma classe, dentre um conjunto de classes pré-definidas, com base nos valores de seus atributos. Esta metodologia ficou composta por técnicas para a análise exploratória dos dados e por técnicas de DM, Regressão Logística Binária (RLB), da Geração de uma Superfície que Minimiza Erros (GSME-PL), da Função Discriminante Linear de Fisher (FDLF), das Redes Neurais Artificiais (RNAs) e da Máquina de Vetores de Suporte (SVM), que foram utilizadas comparativamente, visando a maximização da acurácia quanto à classificação de instâncias.

1.5 ESTRUTURA DO TRABALHO

O trabalho está organizado em seis capítulos, conforme descrito a seguir, incluindo este primeiro capítulo.

O segundo capítulo apresenta a descrição dos problemas aqui abordados que ilustram a utilização da metodologia aqui proposta.

O terceiro capítulo contempla a revisão bibliográfica dos temas relacionados com KDD e DM, além de trabalhos correlatos ao tema aqui abordado.

O quarto capítulo apresenta a metodologia preliminar, ou seja, a proposta para se abordar problemas para a Reconhecimento de Padrões no que diz respeito a tarefa de classificação.

O quinto capítulo apresenta os resultados preliminares, tanto da análise exploratória dos dados, que envolve o teste T^2 de Hotelling, quanto das técnicas de DM aqui analisadas RLB, GSME-PL, FDLF, RNAs e SVM. São propostos alguns testes com o intuito de se obter resultados com a máxima acurácia.

O sexto e último capítulo apresenta as considerações finais da presente pesquisa.

2 DESCRIÇÃO DO PROBLEMA

Neste capítulo são descritos os problemas aqui abordados por meio da metodologia proposta: de uma instituição de ensino superior, de diagnóstico médico e de calibração de balança. Além da descrição são, também, apresentados os conceitos básicos relacionados aos dois problemas.

2.1 INSTITUIÇÃO DE ENSINO SUPERIOR – PÓS GRADUAÇÃO LATO SENSU

As Instituições de Ensino Superior (IES) vêm sofrendo pressões do mercado, por um aperfeiçoamento de seus cursos de PG Lato Sensu, ou seja, em nível de Especialização, ocasionando a concorrência entre as mesmas.

O aumento significativo no número de cursos de PG tem exigido que as instituições se empenhem para alcançar níveis cada vez mais elevados de eficiência em sua gestão, tendo como objetivo fundamental a satisfação de seus alunos. A fim de atender as demandas dos cursos, as IES têm buscado por informações sobre os fatores que atraem estudantes para seus cursos.

É importante ressaltar que o aumento no número de egressos nas IES está diretamente relacionado ao mercado de trabalho, já que os profissionais sofrem constantes pressões para se atualizarem, buscando novas soluções para os diversos novos problemas que surgem em suas organizações diariamente (MAINARDES; DOMINGUES, 2010).

Desta maneira, as perspectivas dos egressos dos cursos podem auxiliar as IES a aprimorarem suas técnicas de ensino, visto a importância de se entender o público a que se destinam, pois, os mesmos desfrutam de uma gama elevada de alternativas e a satisfação está diretamente relacionada à qualidade do serviço prestado.

As técnicas para avaliação dos cursos de PG no Brasil têm sido frequentemente atualizadas, valorizando a coerência entre a proposta do programa e as exigências do Ministério da Educação e Cultura (MEC, 2016). Dentre as referidas exigências, pode-se destacar que o corpo docente deverá ser constituído necessariamente por, pelo menos, 50% de professores portadores de título de Mestre ou de Doutor, obtidos em programa de PG stricto sensu reconhecidos. Os

demais docentes devem possuir, no mínimo, formação em nível de especialização (MECNE, 2016).

Os cursos devem ter duração mínima de 360 horas, sendo que nestas não é computado o tempo de estudo individual ou em grupo, sem assistência docente, e o reservado, obrigatoriamente, para elaboração de Monografia. A duração poderá ser ampliada de acordo com o projeto pedagógico do curso e o seu objeto específico (MEC, 2016).

Para o presente estudo foi realizada a análise de 68 cursos de PG de uma IES privada, cada um deles com 13 módulos. As variáveis independentes (atributos) deste problema, utilizadas foram em um total de 12: “domínio do conteúdo” pelo docente; “didática e clareza na condução do módulo”; “capacidade de despertar a motivação”; “aderência do conteúdo à proposta do curso”; “relacionamento do professor com os alunos”; “planejamento e organização geral”; “sala de aula”; “Eureka & intranet”, onde Eureka é um sistema computacional utilizado para inserir arquivos e outros; “estrutura cantinas e banheiros”; “tutor”; “supervisão acadêmica”; “coordenação do curso” e uma variável dependente (resposta), referente ao índice de satisfação (alunos “satisfeitos” ou “insatisfeitos”).

Foram coletados 885 dados, por meio de um questionário (ANEXO A), sendo que cada um deles é a média de 25 alunos por módulo (total 1.626 alunos). Cada uma destas 885 instâncias (APÊNDICE A) ficou com 12 atributos definidos pelas respostas dos questionários, dos quais 322 ficaram enquadrados na classe “satisfeitos” e 563 apresentaram na classe “insatisfeitos”.

2.2 DIAGNÓSTICO MÉDICO

Com o avanço tecnológico na área medica, novos procedimentos surgem a cada dia, a análise exploratória dos dados e o uso de ferramentas, podem se tornar um fator crucial, em uma tentativa de otimizar todo o processo do diagnóstico e minimizar os riscos, por outro lado, maximizando a eficácia nos resultados.

Existem vários estudos sobre a colestase, divido a frequente incidência tanto na prática clínica de crianças, quanto na de adultos. A colestase é um estado patológico em que há diminuição da formação de bile ou perturbação do seu fluxo. O fluxo biliar pode estar comprometido em qualquer ponto entre os hepatócitos e o duodeno. Embora a bile não esteja fluindo, o fígado continua a produzir bilirrubina,

que “escapa” para o interior da corrente sanguínea (PAULI-MAGNUS; MEIER, 2006; ANTHERIEU *et al*., 2013).

A bilirrubina é então depositada na pele, a qual é parcialmente reabsorvida no intestino e excretada pela urina, causando icterícia (cor amarelada da pele). A icterícia ocorre pelo acúmulo no sangue de bilirrubina direta ou indireta. O acúmulo da bilirrubina direta deve-se a uma colestase (acumulação de bile), por algum impedimento do fluxo natural da bile do fígado ao intestino pelo colédoco (VAN DE STEEG *et al*., 2012).

As causas de colestase são divididas em dois grupos: as intra-hepáticas (originadas no interior do fígado) e as extra-hepáticas (originadas fora do fígado). A colestase intra-hepático é o transtorno no fluxo da bile devido à lesão nos hepatócitos, canalículos biliares, ou ductos biliares intra-hepáticos (PAULI-MAGNUS; MEIER; STIEGER, 2010).

Já a colestase extra-hepático é a alteração do fluxo biliar através dos grandes ductos biliares por obstrução mecânica ou constrição devido a processos benignos ou malignos. As causas extra-hepáticas incluem o cálculo no interior do ducto biliar, a estenose (estreitamento) do ducto biliar, o câncer no ducto biliar, o câncer de pâncreas e a inflamação do pâncreas (ROEB *et al*., 2003; KRISHNAMURTHY; KRISHNAMURTHY, 2009). Somente o cálculo e o câncer no ducto biliar serão considerados no presente trabalho fazendo uso da metodologia proposta. Alternativamente, tais problemas poderiam ser analisados através de exames como a ultrassonografia e, eventualmente, tomografia axial computadorizada.

As variáveis independentes (atributos) utilizadas nesse estudo foram em um total de 14 oriundas de medidas de exames clínicos sugeridos por médico especialista da área: Idade, Sexo, Bilirrubina total, Bilirrubina direta, Bilirrubina indireta, Fosfatases alcalinas, SGOT (Transaminase Glutâmico-Oxalacética), SGPT (Transaminases Glutâmico-Pirúvicas), Tempo de atividade da protrombina, Albumina, Amilase, Creatinina, Leucócitos e Volume Globular, além da variável dependente (resposta classificatória: câncer ou cálculo no duto biliar). Tais variáveis foram obtidas de 118 pacientes (APÊNDICE B) do Hospital das Clínicas (HC) de Curitiba, PR, dos quais, comprovadamente, 35 possuíam câncer e 83 possuíam cálculo no duto biliar (STEINER *et al.*, 2006). Todo este levantamento teve o acompanhamento de um médico especialista da área.

2.3 EMPRESA DE BALANÇAS

Mudanças cada vez mais rápidas e significativas no ambiente competitivo demandam uma incessante busca por qualidade e produtividade. Desta maneira, as indústrias investem continuamente em seus profissionais, em tecnologia e infraestrutura para atender às reais necessidades de todos os clientes, destacando-se as empresas de calibração de balança. É nesse contexto que a confiabilidade adquire um elevado grau de importância, dado o seu enorme potencial para o aumento de produtividade e melhoria de qualidade dos serviços.

A balança é um equipamento que mede a massa de um corpo e a sua calibração contribui para que ela opere dentro das especificações metrológicas, assegurando a confiabilidade nas pesagens e ganhos reais no processo produtivo da empresa. Segundo a Inmetro (2016), a calibração é um conjunto de operações que estabelece, sob condições especificadas, a relação entre os valores indicados por um instrumento de medição (ou sistema de medição ou valores representados por uma medida materializada ou um material de referência), e os valores correspondentes das grandezas estabelecidos por padrões VIM (Vocabulário Internacional de Termos Fundamentais e Gerais de Metrologia). Além disto, a calibração dos equipamentos garante a qualidade da fabricação de um determinado produto e assegura que os instrumentos usados para controlar o seu produto estão dentro de um critério aceitável.

Desta forma, a calibração deve ser realizada com base em referências técnicas, tais como normas nacional e internacional, documentos orientativos do Inmetro (2016) e, assim sendo, é importante que os equipamentos sejam calibrados por uma empresa acreditada. A calibração realizada por estas empresas garante que o equipamento será avaliado conforme os requisitos da Norma ABNT NBR ISO/IEC 17025:2005 e Norma ABNT ISO 9001:2008, seguindo-se assim todos os controles de documentos, registros, rastreabilidade e itens e cálculos a serem avaliados durante a verificação do equipamento.

O resultado de uma calibração deve ser apresentado em um documento técnico, chamado de “certificado de calibração”, porém possuir um certificado por si só não é suficiente. O certificado de calibração não deve ser apenas uma evidência para o auditor, pois avaliar o desempenho de um equipamento de medição ao longo

do tempo é fundamental para a confiabilidade dos resultados de medição e na obtenção da qualidade de produtos e processos.

O presente estudo foi realizado em uma empresa de balanças, a qual comercializa e também oferece serviços de assistência técnica, manutenção, locação e calibração de balanças na cidade de Curitiba – PR.

Foram coletados 1540 dados (APÊNDICE C) por meio de registros de fichas cadastrais do serviço de calibração de balanças referentes aos anos de 2014 e 2015, por meio de um técnico, devidamente treinado e capacitado, que verifica erros e leituras incorretas do equipamento a partir de métodos de ensaio que verificam a exatidão, a precisão e o posicionamento da carga em diferentes regiões da balança.

As variáveis independentes (atributos) utilizadas neste caso foram em um total de 14: Mês; Cliente; Técnico; Fabricante; Capacidade; Corrente de ar; Vibração; Local da Calibração; Temperatura Inicial; Variação da Temperatura; Umidade Relativa Inicial; Variação da Umidade Relativa; Pressão Atmosférica; Variação da Pressão Atmosférica e uma variável dependente (resposta), referente a qualidade da calibração (“boa qualidade” ou “baixa qualidade”).

3 REVISÃO DA LITERATURA

A fundamentação teórica do presente trabalho compreende conceitos gerais a respeito de KDD, cuja a principal etapa é o DM. São aqui abordadas três técnicas para a análise exploratória dos dados: análise descritiva dos dados; detecção de *outliers* e o teste T^2 de Hotelling e também cinco técnicas de DM: RLB, GSME-PL, FDLF, RNAs e SVM, comparativamente. São apresentadas, ainda, as características específicas de trabalhos correlatos recentes, apresentados por pesquisadores das mais diversas áreas, dos quais detalha-se o problema em si, as técnicas utilizadas para a sua resolução, assim como os resultados obtidos.

3.1 O PROCESSO KDD E DATA MINING

A capacidade de gerar e armazenar dados tem aumentado consideravelmente no decorrer dos últimos anos, gerando amplas bases de dados nos mais diversos ramos de atividades. Esse crescimento na quantidade de dados armazenados, por sua vez, tem gerado a necessidade por novas técnicas e ferramentas que possam auxiliar na transformação desses dados em informação útil e conhecimento.

Nesse contexto, o processo de KDD tem sido utilizado para extrair conhecimento de grandes bases de dados e tem se mostrado efetivo na gestão de informações, buscando identificar as mais relevantes e transformá-las em conhecimento útil à tomada de decisão.

Fayyad, Piatetsky-Shapiro e Smith (1996b) definem KDD como um processo de várias etapas não trivial, de extração de informações implícitas, interativo (o usuário pode intervir e controlar o curso das atividades) e iterativo (sequência finita de operações onde o resultado de cada etapa é dependente dos resultados das etapas que as precedem). Tal processo é utilizado para a identificação de padrões acessíveis, válidos, inéditos e potencialmente úteis. Por ser um processo exploratório, é importante que o KDD produza respostas rápidas. Porém, devido à ampla quantidade de dados e ao alto custo computacional dos algoritmos na extração de conhecimento, isto nem sempre é possível. Desta forma, esta ocorrência torna muitas vezes uma resposta aproximada e rápida mais interessante do que uma resposta exata e demorada.

Outro fator importante do KDD é o tipo de dados a serem analisados. A maior parte dos algoritmos de extração de conhecimento estabelece que os dados a serem analisados não tenham muitos atributos, estejam limpos e sem ruídos.

Desta forma, Fang e Rachamadugu (2009) definem o processo KDD como uma evolução natural da tecnologia de informação. Atualmente é possível buscar e armazenar amplas quantidades de dados, todavia, essa quantidade de dados gerados, em geral, supera a capacidade humana de compreensão dos mesmos, o que cria a necessidade de ferramentas para realizar esta análise.

O processo KDD possui forte relação com o aprendizado de máquina, com o Reconhecimento de Padrões (RP), com a Estatística e com a Visualização de dados, objetivando descobrir os padrões nos dados (XIAO; FAN, 2014). O KDD pode ser visto como a confluência dessa ciência (PADHY; MISHRA; PANIGRAHI, 2012). A Estatística proporciona métodos de quantificação da incerteza inerente, bem como procura inferir padrões gerais a partir de amostras de uma população. As técnicas de Visualização de Dados estimulam naturalmente a percepção e a inteligência, aumentando a capacidade de entendimento e de associação de novos padrões (REZENDE *et al.*, 2003). Assim, é preciso entender todas as etapas do KDD e saber como aplicá-las para obter os melhores resultados possíveis.

Praticamente todas as áreas de conhecimento podem usar o processo KDD. Maimon e Rokach (2010) citam algumas destas áreas:

- *Marketing*: análise comportamento do cliente; identificar diferentes grupos de clientes;
- Medicina: sintomas de doenças, análise de experimentos, efeitos colaterais de medicamentos;
- Detecção de fraude: monitoramento de crédito, chamadas clonadas de telefones celulares, identificar transações fraudulentas;
- Agricultura: tendências e classificação de pragas em legumes e frutas;
- Área social: pesquisa de intenção de votos, resultados de eleições;
- Militar: análise de informações; perfil de usuários de drogas;
- Ciência espacial: astronomia, análise de dados espaciais.

3.1.1 Etapas do KDD

As etapas do KDD contêm uma sequência de passos que auxiliam nas mais variadas decisões a serem tomadas. Toda fase possui uma interseção com as demais, melhorando assim a cada resultado (RELICH; MUSZYNSKI, 2014). O processo KDD é composto por cinco etapas, a seguir:

- Seleção dos dados;
- Pré-processamento dos dados;
- Transformação dos dados;
- Mineração de dados e
- Interpretação e avaliação dos resultados.

Estas etapas estão ilustradas na Figura 3.1 a seguir, representando suas várias fases.

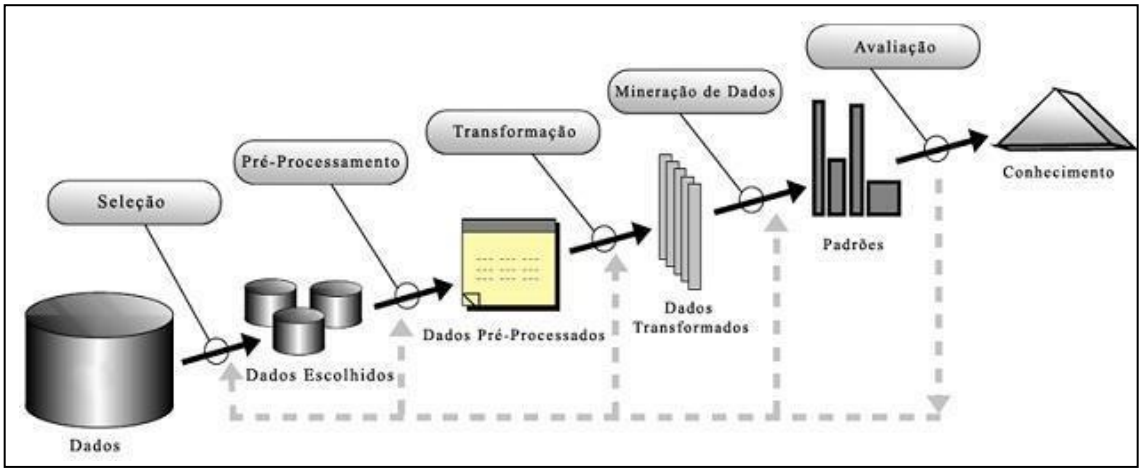

Figura 3.1 - Etapas do Processo KDD**.**
Fonte: Fayyad, Shapiro e Smyth (1996c).

O objetivo das diversas etapas do processo KDD é fazer com que os dados brutos se tornem conhecimento útil. Tais etapas são representadas por um conjunto de fases orientadas por suas atividades que são direcionadas à extração, manipulação e DM.

Segundo Barros e Campos (2006), as etapas do KDD, que antecedem a DM, podem levar até 80% do tempo necessário para todo o processo de análise devido

às dificuldades de integração de bases de dados com estruturas variadas. A Figura 3.2 ilustra o tempo necessário para cada etapa do processo KDD.

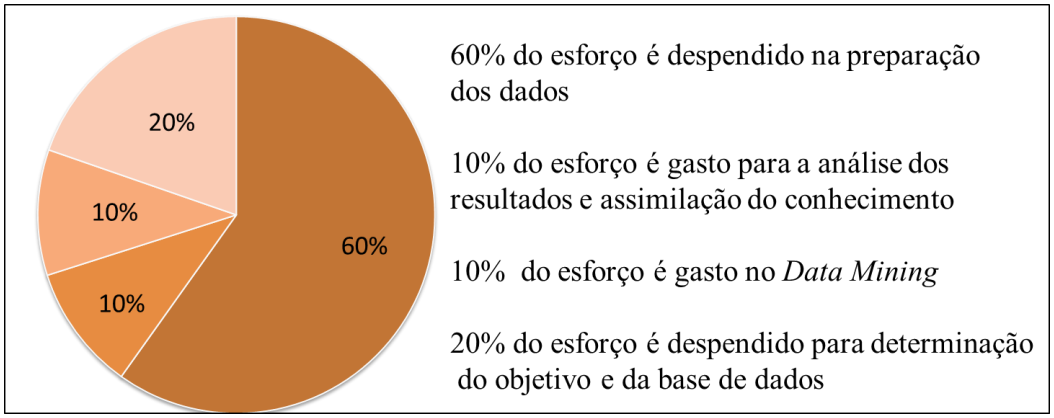

Figura 3.2 - Esforço referido para cada etapa do processo KDD.
Fonte: Adaptado de Cabena et al. (1998) e Sassi (2006).

Conforme a Figura 3.2, observa-se que 60% do tempo necessário é para a preparação dos dados (seleção, pré-processamento e transformação), com o objetivo de organizar os dados para facilitar as tarefas realizadas pelas etapas posteriores do processo KDD obtendo-se, consequentemente, resultados com maior qualidade. No desenvolvimento deste trabalho é utilizado o processo KDD definido por Relich e Muszynski (2014) e, assim sendo, as etapas são explicitadas a seguir:

a) Seleção de Dados

Esta etapa é também conhecida como “Redução de Dados”. É a primeira etapa no processo de descoberta de informação e possui papel fundamental no resultado final, uma vez que nesta etapa é definido o conjunto de dados contendo todas as possíveis variáveis (atributos) e registros (instâncias ou casos ou observações ou padrões) que se pretende analisar. Em sua grande maioria, esta seleção é realizada por um especialista da área, ou seja, alguém que realmente entende do assunto em questão.

O processo de seleção é bastante complexo, já que os dados podem ser selecionados das mais diversas fontes, tais como: banco de dados relacional, arquivo texto legado, *Data Warehouses* (Armazéns de Dados), planilhas, dentre outros. Estes dados são separados de acordo com a necessidade e objetivo do projeto, sendo comum a necessidade de se utilizar um *software* específico para a realização desta seleção.

b) Pré-processamento e Limpeza dos Dados

Nesta etapa são realizadas tarefas que excluem dados redundantes e inconsistentes, recuperam dados incompletos e avaliam possíveis discrepâncias nos dados. O auxílio do especialista do domínio é fundamental, pois é o mesmo que definirá se os atributos adquiridos são interessantes, se o conhecimento é válido, novo e útil, ou se será necessário retornar a alguma das etapas anteriores (ANUMALLA, 2007).

Além disto, nesta etapa é verificada a possibilidade de diminuir o número de variáveis envolvidas no processo, compreende a identificação de quais informações, dentre as bases de dados existentes, devem ser analisadas durante o processo KDD, visando melhorar o desempenho dos algoritmos de análise (GOLDSCHMIDT; PASSOS, 2005; PADHY; MISHRA; PANIGRAHI, 2012). Para isto, podem ser aplicados métodos estatísticos, a fim de melhorar a eficácia dos algoritmos de classificação, como apresentado por Steiner *et al*. (2006).

Segundo Zhang, Zhang e Yang (2003), essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido as bem conhecidas dificuldades de integração de bases de dados heterogêneas. Além disso, podem aparecer problemas que são específicos para cada aplicação e que precisarão ser resolvidos com soluções específicas. É recomendado selecionar algumas amostras randomicamente, a fim de obter uma ideia do que pode ser esperado.

A limpeza dos dados envolve a verificação de ruídos, dados estranhos ou inconsistentes, e nesta etapa são estabelecidas as estratégias para resolver os problemas de ausência de dados. É realizada através de um pré-processamento dos dados, visando adequá-los aos algoritmos. Isto ocorre por meio da integração de dados heterogêneos, tratamento de ausências de dados, eliminação de dados incompletos, repetição de registros, de dados estranhos e/ou inconsistentes. E uma “boa limpeza dados” é essencial, podendo inclusive diminuir o tempo de processamento, eliminando consultas desnecessárias à base de dados.

c) Transformação dos Dados

A etapa de transformação dos dados ou codificação dos dados tem como objetivo principal converter o conjunto bruto de dados em uma forma padrão de uso (GOLDSCHMIDT e PASSOS, 2005). Esta etapa é implementada através de um

processamento dos dados, visando organizar os dados para auxiliar o trabalho sucedido pelas fases posteriores do processo KDD.

Esta transformação dos dados se refere a uma transformação que seja aplicada a todos os valores de um determinado atributo para todos os atributos. No entanto, não existe um critério único de transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos.

A normalização dos dados, por exemplo, é uma transformação, que consiste em ajustar a escala dos valores de cada atributo de maneira que os valores permaneçam em pequenos intervalos, tais como: [-1, 1] ou [0, 1]. O agrupamento é um processo de organização de elementos de um conjunto em grupos, cujos membros são similares em alguma característica, por meio de uma grande quantidade de atributos com tamanhos diferentes. A suavização remove valores “errados” dos dados. A generalização converte valores muito específicos para valores genéricos. Novos atributos podem ser gerados a partir de outros já existentes (HAN; KAMBER; PEI, 2011). Tais técnicas se fazem necessárias para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de maneira tendenciosa deliberados métodos de DM.

Além disto, a etapa de transformação dos dados considera a unificação e a concentração dos dados selecionados e limpos nas fases antecedentes, de modo a diminuir o tempo de processamento dos mecanismos de mineração e facilitar os algoritmos quanto a acurácia dos resultados. Uma desvantagem da etapa de transformação é a redução da medida de qualidade do conhecimento a ser descoberto, podendo-se perder alguns detalhes relevantes.

d) Mineração dos Dados

Segundo Tan, Steinbach e Kumar (2009) é importante não confundir DM e KDD, visto que KDD é todo o processo até que se chegue ao resultado de um padrão de comportamento das variáveis.

Para Larose (2005), “DM é o processo de descobrir novas correlações significativas, atributos e tendências peneirando grandes quantidades de dados armazenados em repositórios, usando as tecnologias de RP”.

Saitta, Raphael e Smith (2005) definem DM como uma análise de grandes conjuntos de dados a fim de descobrir relacionamentos inesperados e de resumir os

dados de uma forma que eles sejam úteis, assim como compreensíveis ao empregador dos dados.

Xiao e Fan (2014) afirmam que DM é dada por meio de um campo interdisciplinar que reuni técnicas de máquinas de conhecimentos, RP, estatísticas, banco de dados e visualização, que consigam extrair informações de amplas bases de dados.

Choudhary, Harding e Tiwari (2009) associam a DM como uma etapa no processo KDD, como já mostrado, que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, determinam um conjunto de atributos de exatos dados.

Witten, Frank e Hall (2011) consideram que as técnicas de DM se referem a extração automatizada ou conveniente de padrões, representando conhecimento implicitamente armazenado em bases de dados amplas, armazéns de dados, e outros repositórios de informação de grande porte.

Assim sendo, a principal característica do DM é o algoritmo minerador, que diante de uma tarefa específica será capaz de extrair de maneira competente, o conhecimento implícito e útil de um banco de dados, podendo auxiliar na previsão de um conhecimento futuro.

DM é uma área emergente dentro da inteligência computacional usada na análise de grandes bancos de dados, com a geração de atributos e a extração de informações dessas bases, podendo, por exemplo, examinar as relações de similaridade entre as informações (GURULER; ISTANBULLU; KARAHASAN, 2010).

Desta forma, DM consiste da utilização de técnicas de RP, aprendizado de máquina e estatística, para a classificação, predição, agrupamento, sumarização, modelos de relacionamento entre variáveis, modelo de dependência, análise de séries temporais ou associação de atributos. As técnicas são concebidas para agir sobre grandes bancos de dados, com o intuito de descobrir atributos úteis e recentes que poderiam de outra forma, permanecer ignorados. Estas técnicas vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como Redes Neurais, Lógica Difusa e Algoritmos Genéticos (AGs). Por isso, é considerada uma das áreas interdisciplinares mais promissoras em desenvolvimento na tecnologia de informação.

DM se dá por meio do processo de extração de conhecimento, mas apesar de encontrarmos várias ferramentas que nos auxiliam na execução dos algoritmos de mineração, os resultados muitas vezes deixam a desejar, necessitando de uma análise humana. Todavia, ainda assim, a mineração colabora de forma expressiva no processo do KDD, permitindo aos especialistas concentrarem esforços somente em partes mais significativa dos dados.

d) Interpretação e Avaliação dos Resultados

Em seguida à etapa de DM, é necessária a interpretação e avaliação do resultado obtido, última etapa do processo KDD. Esta etapa também é conhecida como pós-processamento e envolve todos os participantes que avaliam de forma criteriosa os resultados proporcionando uma interpretação para o modelo, de onde se extrai o conhecimento. Caso o resultado não seja satisfatório, o processo pode retornar a qualquer uma das etapas anteriores.

Essa interpretação deve ser incluída no algoritmo minerador, mas determinadas vezes é proveitosa a implementação separadamente. Desta forma, a principal finalidade dessa etapa é garantir um bom grau de compreensão do conhecimento descoberto pelo algoritmo minerador, validando-o por meio de medidas da qualidade da solução e da percepção de um analista de dados. Esses dados serão consolidados em forma de relatórios demonstrativos com a documentação e explicação das informações relevantes ocorridas em cada etapa do processo de KDD. Uma maneira genérica de alcançar o entendimento e interpretação dos resultados é empregar técnicas de visualização (NGAI *et al*., 2011).

As técnicas de visualização estimulam a percepção e a inteligência humana, desenvolvem a capacidade de entendimento e a associação de novos atributos (WITTEN; FRANK; HALL, 2011).

Em geral, a avaliação final do processo deve ser compreensível. Todavia, definir a habilidade para compreender não é uma tarefa trivial. Em certos contextos, a compreensão pode ser medida pela simplicidade do modelo (como, por exemplo, número de nós de uma árvore de decisão). Todavia, até o momento, não existe um mecanismo efetivo para medir a aptidão para compreender o conhecimento.

A compreensão é útil para validar o conhecimento, para descoberta de novos atributos, para a sugestão de melhores atributos e para o aprimoramento do conhecimento.

3.1.2 Tarefas do KDD

O processo KDD é capaz de realizar várias tarefas. Para Han, Kamber e Pei (2011) não há uma tarefa que atenda a todos as condições e, por este motivo, é importante conhecer suas características para conduzir os dados a serem minerados para uma tarefa específica ou para um conjunto de tarefas. Desse modo, a aplicação dos dados em uma tarefa de forma errada contribui para descobertas irrelevantes no processo de extração de conhecimento.

As tarefas do KDD funcionam com base nas técnicas de DM. Estas técnicas podem ser abordadas na forma de preditiva e descritiva (HAN; KAMBER; PEI, 2011). De acordo com Rud (2001) e Cios *et al*. (2007), as tarefas podem ser divididas em aprendizagem supervisionada (tenta explicar ou categorizar dados em particular) e aprendizagem não-supervisionada (tenta encontrar atributos ou similaridades entre grupos de registros sem o uso de um campo em particular como alvo ou de conjuntos de classes pré-definidos, ou seja, sem categorizar).

As tarefas mais comuns, segundo Relich e Muszynski (2014), são associação, classificação, regressão, *clustering* e visualização de dados, conforme ilustradas na figura 3.3.

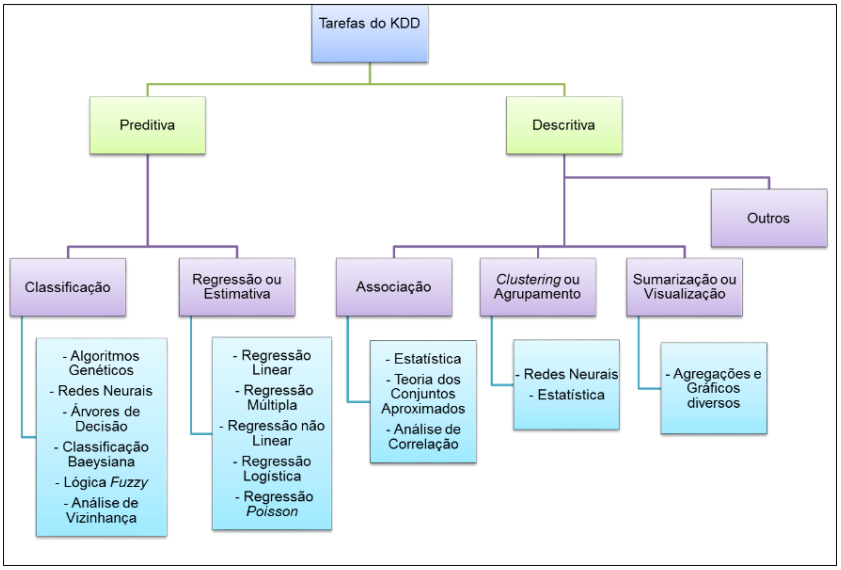

Figura 3.3 - Algumas tarefas do processo KDD e suas técnicas na MD.
Fonte: Adaptado de Maimon e Rokach (2010) e Rezende (2005).

Observa- se que cada tarefa do processo KDD possui técnicas diferentes associadas a DM e as mais conhecidas, segundo Witten, Frank e Hall (2011), são: Árvores de Decisão, Regras de Classificação, Redes Neurais, Vizinhos Mais Próximos, Regressão Linear ou Não Linear e AGs. Existem ainda abordagens híbridas, que aplicam duas ou mais técnicas em conjunto. Não se pode dizer que existe uma melhor técnica, já que o desempenho de cada uma delas dependerá do problema abordado.

3.1.2.1 Classificação

De acordo com Carvalho (2001) “a classificação é uma das técnicas mais utilizadas no DM simplesmente porque é uma das tarefas cognitivas humanas mais realizadas no auxílio à compreensão do ambiente em que vivemos”.

Segundo Tan, Steinbach e Kumar (2009), a classificação pode ser considerada uma tarefa mal definida, não determinística, que é inevitável pelo fato de envolver predição. A variável de predição poderá ser discreta ou categórica.

Para Kamsu-Foguem, Rigal e Mauget (2013), a classificação é a tarefa mais estudada no processo KDD e tem como objetivo encontrar um conhecimento que possa ser empregado para prever a classe de um determinado registro. Assim sendo, a classificação procura estudar um conjunto de registros históricos (atributos) e elaborar descrições de suas características em cada uma das classes.

Yoo *et al*. (2012) definem a classificação como sendo uma aplicação de um conjunto de exemplos pré-classificados, para desenvolver um modelo capaz de classificar uma população maior de registros. A classificação cria automaticamente um modelo através de um conjunto inicial de registros. O modelo é composto de instâncias (padrões; exemplos; dados), os quais são usados para diferenciar as classes e quando alcançada, este é usado para classificar automaticamente os demais registros (MA, 2012).

Em geral, a classificação consiste em construir um modelo que classifique um item ou registro de dados em uma classe dentre algumas pré-definidas. Cada classe corresponde a um padrão único de valores dos atributos, sendo que este padrão único pode ser considerado a descrição da classe. O objetivo da classificação é encontrar algum tipo de relação entre os atributos e as classes, de modo que o

processo de classificação possa usar esse relacionamento para prever a classe de um exemplo novo e desconhecido (SANTOS; AZEVEDO, 2005).

Os métodos de classificação têm aplicação em diversas áreas tais como análise de crédito; ações de *marketing*; detecção de fraudes; telecomunicações; segmentação de clientes; modelagem de negócios; e em diagnósticos médicos (NGAI *et al*., 2011). Por exemplo, alguma pessoa pode classificar doenças e ajudar a prever tipos de doenças baseados nos sintomas dos pacientes.

3.1.2.2 Regressão

Para Fayyad, Piatetsky-Shapiro e Smith (1996b), a regressão consiste em descobrir uma função que represente um item de dados para uma variável de predição de valor numérico contínuo.

Segundo Alzghoul, Löfstrand e Backe (2012), a regressão trata de “aprender” uma função que mapea um item de dado para uma variável de predição real estimada. Pode ser utilizada para executar uma tarefa de classificação, convencionando-se que diferentes faixas de valores contínuos correspondem a diferentes classes.

A regressão é utilizada para definir um valor para alguma variável contínua desconhecida como, por exemplo, receita, altura, ou saldo de cartão de crédito; estimar a renda total de uma família; estimar o valor em tempo de vida de um cliente; estimar a probabilidade de um paciente morrer baseando-se nos resultados de um conjunto de diagnósticos médicos (HAN; KAMBER; PEI, 2011; VIAENE *et al*., 2007).

3.1.2.3 Associação

Datta *et al.* (2006) definem que a tarefa de associação pode ser considerada como uma tarefa bem definida, determinística e relativamente simples, que não envolve predição da mesma forma que a tarefa de classificação.

A tarefa de associação permite relacionar a ocorrência de um determinado conjunto de itens com a ocorrência de outro conjunto de itens. Segundo Yoo *et al*. (2012), regras de associações buscam identificar semelhanças entre registros de um subconjunto de dados, sendo que essas semelhanças/associações são expressas na forma de regras.

A associação está relacionada à descoberta de regras de associação ou correlação indicando condições de atributo-valor que ocorrem frequentemente juntas em um conjunto de dados. Esta estratégia é geralmente utilizada em aplicações onde se busca identificar itens que possam ser colocados juntos em um mesmo tipo de negociação ou, ainda, pode ser utilizada para avaliar a existência de algum tipo de relação temporal entre os itens constantes de uma base de dados.

Desta maneira, para definir o quanto esta regra é representativa para o conjunto de dados utilizam-se dois fatores conhecidos como suporte e confiança. O suporte revela a probabilidade dos objetos da base de dados possuírem os atributos envolvidos na regra, enquanto que a confiança revela a proporção de objetos que possuem os atributos do antecedente e do consequente. Tanto o suporte quanto a confiança funcionam como uma classe de filtro para a obtenção das regras e, geralmente, o próprio usuário que os define (GRANATYR, 2011; MARTÍNEZ-DE-PISÓN *et al.*, 2012).

Existem diversos algoritmos consagrados para DM por meio de regras de associação, dentre eles destacamos: *Apriori*, *Partition* e *Multiple Level*, onde o primeiro é o mais utilizado e os demais, ou são extensões deste ou o utilizam. Maimon e Rokach (2010) apontam como principais algoritmos de associação os seguintes:

- Algoritmo *Apriori*: responsável por descobrir o conjunto de itens frequentes por meio de múltiplos passos executados na base de dados iniciando com um conjunto semente de itens que determinará novos conjuntos potenciais, chamados de conjunto de itens candidatos;
- Algoritmo *Basic*: difere do algoritmo *Apriori* nas transações. O algoritmo cria novas transações conhecidas como transações estendidas onde são adicionados todos os itens antecessores de um dado item;
- Algoritmo *Cumulate*: este algoritmo pré-computa os antecessores de um item e se estes itens forem relevantes, os adiciona na transação.

Estes métodos são usados, por exemplo, para criar regras de associação, que podem ser usadas em uma análise de cesta básica ou em cestas de compras em que produtos são levados juntos pelos consumidores, personalização de páginas da *Web*, construção de catálogos, definição de promoções, entre outros.

3.1.2.4 *Clusterização* ou Agrupamento

A palavra *clusterização* é um neologismo do termo *clustering*, que difere da classificação, já que a primeira, visa criar os agrupamentos por meio da organização dos elementos, enquanto que a segunda, busca alocar elementos em classes já pré-definidas (GUELPELI, 2009).

Segundo Harrison (1998) e Hastie, Tibshirani e Friedman (2009) trata-se de um processo de partição de uma população heterogênea em vários subgrupos ou *clusters* mais homogêneos, com o objetivo de formar agrupamentos ótimos sobre os dados, dividindo iterativamente o conjunto de exemplos em k-partições mutuamente exclusivos, as quais devem maximizar uma função critério pré-definida (BERKHIN, 2002).

Para Xiao e Fan (2014), agrupamento é uma tarefa onde se busca identificar um conjunto finito de categorias ou agrupamentos para descrever os dados. O agrupamento assegura que a similaridade inter-segmentos seja baixa, enquanto que a similaridade intra-segmentos seja alta (SANTOS; AZEVEDO, 2005).

Desta forma, a *clusterização* identifica possíveis agrupamentos nos dados, tal que os elementos de uma classe tenham alta similaridade entre si e sejam muito diferentes dos elementos das outras classes. Por exemplo, podem-se agrupar as casas de uma área de acordo com sua categoria, área construída e localização geográfica (HAN; KAMBER; PEI, 2011; SHIUE; GUH; TSENG, 2012).

3.1.2.5 Sumarização

Segundo Fayyad *et al*. (1996a), a tarefa de sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Para Goldschmidt e Passos (2005) consiste em procurar identificar e indicar características comuns entre conjuntos de dados. A tarefa de sumarização deve buscar por características que sejam adequadas a uma parte significativa de clientes.

A sumarização determina uma descrição com dispersão compacta para um dado subconjunto no pré-processamento dos dados, frequentemente empregadas na análise de descobrimento de dados, por exemplo, derivação de regras resumidas ou visualização multivariada.

3.1.3 Técnicas de Análise Exploratória dos dados

Existem na literatura, diversas ferramentas que dão suporte ao processo KDD. As diversas ferramentas de análise dispõem de um método baseado na verificação, isto é, o utilizador constrói hipóteses específicas e posteriormente, verifica ou refuta as mesmas.

Harrison (1998) assegura que não há uma ferramenta que resolva todos os problemas de KDD. Diferentes métodos servem para diferentes propósitos e a familiaridade com as ferramentas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados. A seguir são descritas as ferramentas normalmente usadas no KDD através da DM para este trabalho.

3.1.3.1 Detecção dos *Outliers* (dados atípicos)

- Escore Padronizado ou Escore Z

Os dados atípicos em análises estatísticas são valores excessivamente reduzidos ou elevados, os quais podem distorcer substancialmente os resultados. Desta maneira, sempre que um *outlier* estiver presente nos dados, os resultados gerados com e sem este caso “incomum” podem ser muito diferentes, levando a conclusões conflitantes (RIBAS, 2011).

Sendo assim, a principal razão de se examinar as instâncias, visando à identificação de *outliers*, é a necessidade de se tratar adequadamente tais instâncias. É possível identificar *outliers* univariados, ou seja, valores incomuns dentre as variáveis, verificando-se cada uma delas, por meio de análise descritiva ou inspeção visual do gráfico. Eles apresentam as seguintes propriedades: as magnitudes de seus escores padronizados são maiores do que (3) ou menos do que (-3) e, além disso, não estão integrados aos escores padronizados das instâncias remanescentes. Para identificá-los é recomendável obter, inicialmente, os escores padronizados de todas as variáveis (RIBAS, 2011).

O escore padronizado ou o escore *z* permite a comparação de valores de diferentes conjuntos de dados. É obtido pela conversão de um valor para uma escala padronizada.

O escore padronizado é o número de desvios padrões a que se situa determinado valor de *x*, acima ou abaixo da média, expresso pela equação (3.1), a seguir.

$$Z = \frac{X - \mu}{\sigma} \qquad ou \qquad Z = \frac{X - m}{s} \tag{3.1}$$

onde:

x = escore bruto; μ = média populacional; σ = desvio padrão populacional; *m* = média amostral; *s* = desvio padrão amostral. Um resultado para *z* = 2, por exemplo, significa que o valor está “2 desvios padrões” acima da média; um resultado para *z* = – 3, significa que o valor está a “3 desvios padrões” abaixo da média.

- Distância de *Mahalanobis*

A distância de *Mahalanobis* é a distância entre um dado (instância) e um centróide de espaço multivariado (média geral). É uma métrica que difere da distância Euclidiana por levar em consideração a covariância entre os conjuntos de dados (MINITAB, 2016).

Assim sendo, é um método multivariado utilizado para detectar dados atípicos (*outliers*); este teste é muito utilizado na análise de *clusters* e outras técnicas de classificação. A distância de *Mahalanobis* entre dois vetores de uma mesma distribuição de probabilidade, que possuam uma matriz de covariância *S*, é definida pela equação (3.2).

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})} \tag{3.2}$$

onde:

Σ^{-1} = inversa da matriz de covariância *S*; $\vec{x}, \vec{y}$ = medida de dissimilaridade entre dois vectores aleatórios.

Caso a matriz de covariância seja diagonal, a distância de *Mahalanobis* se reduz à distância Euclidiana normalizada, definida pela equação (3.3).

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1} \frac{(x_i - y_i)^2}{\sigma_i^2}} \tag{3.3}$$

onde:

p = quantidade de vetores; σ_i = desvio padrão do conjunto amostral; x_i e y_i = coordenadas dos elementos que compõem o conjunto amostral.

3.1.3.2 Teste T^2 de Hotelling

Hotelling (1947) foi um dos primeiros a analisar variáveis correlacionadas sob uma perspectiva de controle estatístico, utilizando-se de um procedimento multivariado de controle em dados contendo informações sobre localizações de bombardeios, durante a Segunda Guerra Mundial.

O teste T^2 de Hotelling é aplicado com o intuito de verificar se as duas populações representadas por suas amostras são originadas de populações distintas, ou seja, se existe diferença nas suas várias características médias.

O teste T^2 de Hotelling constrói o elipsóide de confiança que permite verificar se o processo está ou não sob controle, considerando-se todas as características simultaneamente (JOHNSON e WICHERN, 2002). O teste é baseado na normalidade da coordenada tangente e na distribuição normal *p*-variada, e é utilizado em controle estatístico de processos multivariados e a isotropia não é assumida.

Desta forma, o procedimento abordado por Hotelling para avaliação do vetor de médias, permite o controle simultâneo das médias de várias características de qualidade correlacionadas e é mais sensível para detectar desvios que são fracamente detectados por gráficos de controle das variáveis individuais (HE e GRIGORYAN, 2005).

Devido à complexidade dos cálculos que abrange o conhecimento da álgebra matricial, a aceitação dos gráficos de controle multivariados tem sido lenta, embora o gráfico T^2 de Hotelling seja o mais conhecido no controle multivariado do processo.

O gráfico de controle T^2 de Hotelling proporciona mais sensibilidade do que os gráficos univariados, permitindo ao operador detectar mais rapidamente os possíveis problemas existentes no processo e com isso corrigi-los com mais agilidade (HENNING *et al*., 2011). Os gráficos de controle multivariados se baseiam na distribuição T^2 de Hotelling que é a generalização da estatística t de *Student*.

A estatística do teste T^2 de Hotelling é baseada em estimativas amostrais da matriz de covariância e é aplicado para verificar a igualdade dos vetores médios de duas amostras multivariadas A e B, conforme (3.4) e (3.5) a seguir.

$$\frac{T^2\,(m+k-n-1)}{(m+k-2)\,n} \sim F_{n,m+k-n-1} \tag{3.4}$$

onde:

$$T^2 = (X_A + X_B)' [(\frac{1}{m} + \frac{1}{k})\, S_p]^{-1}\, (X_A + X_B) \tag{3.5}$$

O resultado é comparado com a distribuição $F_{n,\ m+k-n-1}$ (0.95) (distribuição *F* de Snedecor a uma probabilidade de 95%). Nas expressões (3.6) e (3.7) têm-se que: *m* e *k* = número de atributos das amostras *A* e *B*, respectivamente; *n* = número de variáveis (atributos) de cada instância; S_p = matriz de covariância conjunta de *A* e *B*; X_A = vetor (*n* x 1) médio da amostra *A*; X_B = vetor (*n* x 1) médio da amostra *B*; S_p^{-1} = inversa da matriz covariância amostral conjunta, sabendo-se que S_p é dada por meio de (3.6).

$$S_p = \frac{(m-1)\,S_A + (k-1)S_B}{m+k-2} \tag{3.6}$$

onde:

S_A = matriz de covariância da amostra *A;* S_B = matriz de covariância da amostra *B*.

Neste teste (3.7), se:

$$\frac{T^2\,(m+k-n-1)}{(m+k-2)\,n} >> F_{n,m+k-n-1}(0{,}95) \tag{3.7}$$

rejeita-se, fortemente, com uma probabilidade de 95%, a hipótese de que as amostras estejam centradas no mesmo vetor de médias.

3.1.3.3 Análise dos Componentes Principais

A Análise dos Componentes Principais (ACP) é uma técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes

principais. Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais; são independentes entre si e são estimadas com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (MORRISON, 1976; KENDALL, 1980; JOHNSON; WICHERN, 2002; VARELLA, 2008; LYRA *et al.,* 2010).

A ACP tem como finalidade representar ou descrever um número de variáveis iniciais a partir de um menor número de variáveis hipotéticas. Isto é, permite identificar novas variáveis, em menor número que o conjunto inicial, mas sem perda significativa da informação contida neste conjunto. Para a determinação dos componentes principais, é necessário calcular a matriz de variância-covariância (Σ) ou a matriz de correlação (R), encontrar os autovalores e os autovetores e, por fim, escrever as combinações lineares que serão as novas variáveis, denominadas de componentes principais. A Figura 3.4 apresenta um esquema de aplicação da ACP (SOUZA, 2000; SOUZA; POPPI, 2012; SILVA; SBRISSIA, 2010).

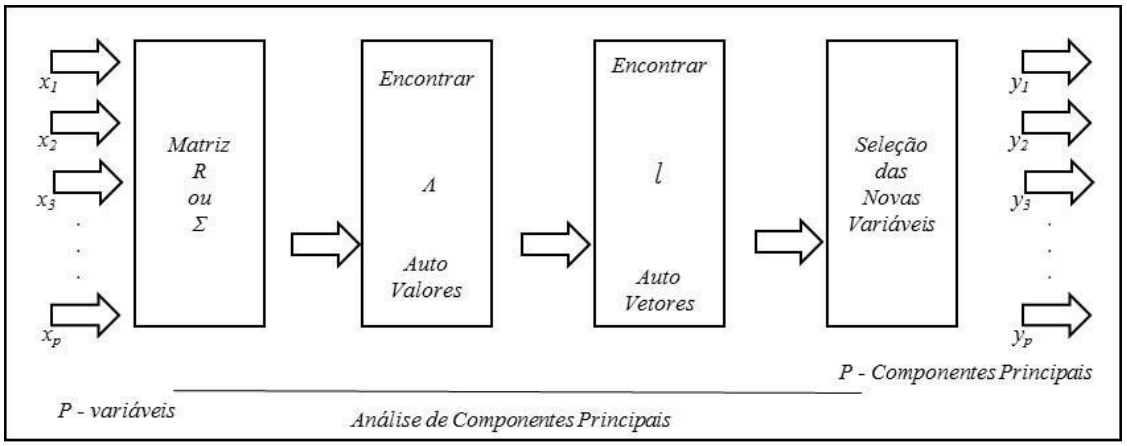

Figura 3.4 - Esquema da aplicação da análise de Componentes Principais
Fonte: Adaptado de Souza (2000)

3.1.4 Técnicas de Data Mining

3.1.4.1 Regressão Logística Binária (RLB)

Segundo Hines (2006), a estatística trabalha com a coleta, apresentação, análise e uso de dados para resolução de problemas, tomada de decisões, desenvolvimento de estimativas e planejamento tanto de produtos quanto de procedimentos, e ainda é usada para a descrição e a compreensão da variabilidade

dos dados. Desta forma, um importante instrumento na estatística é a análise multivariada, que trata todas as variáveis simultaneamente, sumarizando os dados e revelando a sua estrutura com a menor perda de informações possível (JOMBART; PONTIER; DUFOUR, 2009; HAIR *et al*., 2009).

A RLB é definida como uma técnica estatística de análise multivariada que permite o ajuste de um conjunto de variáveis independentes a uma variável de resposta categórica dependente. Ao contrário das variáveis contínuas, as variáveis categóricas podem assumir apenas alguns valores particulares de resposta, podendo estes ser binários (dicotômicos) cuja resposta possui apenas dois níveis (não ou sim) ou politômicos (três ou mais classes), uma extensão da anterior, onde as respostas podem assumir múltiplos níveis de saída (HOSMER; LEMESHOW, 2000; GOUVÊA; GONÇALVES; MANTOVANI, 2012).

A RLB consiste em relacionar, por meio de um modelo, a variável resposta (alunos “satisfeitos” ou “insatisfeitos”) com os atributos que influenciam em sua ocorrência (variáveis pertencentes ao conjunto *A* ou *B*) (HAIR *et al*., 2009).

As premissas básicas a serem atendidas são: a) a média condicional da equação da RLB será um valor definido entre “0” e “1”; b) os erros da equação seguirão a distribuição binária; e c) os resultados obtidos podem ser entendidos na forma de probabilidades (HOSMER; LEMESHOW, 2000).

O modelo de RLB, conforme apresentado por Hosmer e Lemeshow (2000), assume a relação exposta na equação (3.8), também conhecida como função logística. Assim sendo, no modelo de RLB, a variável resposta Y_i é binária, ou seja, poderá assumir dois valores, $Y_i = 0$ ou $Y_i = 1$.

$$Y = \mathrm{f}(x) = (1 + \mathrm{e}-\eta)-1, x \in Rn \qquad (3.8)$$

onde:

η= g(x) é obtido em um ajuste linear. A qualidade do ajuste é medida pela função desvio s_p.

3.1.4.2 Geração de uma Superfície que Minimiza Erros (GSME-PL)

Esta técnica de DM faz uso da Programação Linear (PL). A PL é um modelo matemático constituído de uma função linear (denominada função objetivo), de restrições técnicas e de não negatividade, representadas por um grupo de

inequações ou equações lineares, que sucede das restrições de recursos do problema estudado (BAZARAA; JARVIS; SHERALI, 2009; SILVA *et al*., 2010).

Segundo Rodrigues (2006), a PL está dentro dos métodos de Programação Matemática, onde estes fornecem modelos, na sua maioria determinísticos, normativos e otimizantes, visando problemas de decisão, bem estruturados, cujo o maior desafio é a natureza combinatória das soluções. Deste modo, a PL é uma das técnicas de otimização mais empregadas na solução de problemas de alocação otimizada de recursos, otimização de estratégias, entre outros.

Além disso, os problemas de PL podem ser modelados matematicamente pela necessidade de se otimizar a função objetivo e as restrições lineares, em geral, maximizam os lucros ou minimizam os custos.

Larrosa, Muszinski e Pinto (2011) utilizaram a PL a fim de formular uma pasta a partir de vegetais para produção de sopa desidratada, tendo como objetivo maximizar o valor calórico. Os resultados mostraram que a formulação da pasta de vegetais utilizando a PL foi adequada para maximizar o valor calórico do produto final.

Scaratti e Calvo (2012), por exemplo, desenvolveram um indicador sintético para avaliar a qualidade da gestão municipal da atenção básica à saúde. Para tanto, os autores contemplam simultaneamente os critérios de relevância, de efetividade, de eficácia e de eficiência agregados em medidas de valor, mérito e qualidade. Os resultados desse processo de avaliação foram agrupados em múltiplos critérios de desempenho que refletem a capacidade do gestor municipal de saúde de alocar recursos para atender as necessidades de promoção, prevenção e recuperação da saúde de seus munícipios.

Bennett e Mangasarian (1992), apresentaram o modelo matemático de PL em (3.9), a seguir, que gera um plano que minimiza a média ponderada da soma das violações das instâncias dos conjuntos *A* e *B* que estão do “lado errado” do plano separador. Neste modelo e_k e $e_m \in R^k$ e R^m, respectivamente; *w* é o vetor “peso” $\in R^n$, normal ao plano separador ótimo e $\gamma \in R$, fornece a localização da superfície separadora ótima $wx = \gamma$.

$$\underset{m,\gamma,y,z,}{Min} \quad \frac{e_m y}{m} + \frac{e_k z}{k} \tag{3.9}$$

$$s.a.: A_w - e_m \gamma + y \geq e_m$$

$$- B_w + e_k \gamma + z \geq e_k$$

$$y \geq 0,\ y \in R^m$$

$$z \geq 0,\ z \in R^k$$

3.1.4.3 Função Discriminante Linear de Fisher (FDLF)

A FDLF é uma combinação linear de características originais a qual se caracteriza por produzir separação máxima entre duas classes de objetos ou fixar um novo objeto em uma das duas classes. É uma das técnicas mais usadas quando se trata de dados com estrutura de grupo e também é utilizada para classificar casos de grupos.

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar um determinado elemento (*E*) num determinado grupo de variáveis, entre os diversos grupos existentes π_1, π_2, π_3, ... π_i. A função discriminante constitui em uma combinação linear de variáveis independentes. Pode-se construir uma função discriminante a partir das características de dois grupos de indivíduos e com essa função classificar um novo indivíduo em um dos grupos. Assim sendo, dentro da análise discriminante, um tópico de grande relevância é a FDLF (JOHNSON e WICHERN, 2002).

Deste modo, a ideia de Fisher foi transformar observações multivariadas *X*'s em observações univariadas *Y*'s oriundas das populações π_1 e π_2 de tal modo que estas apresentem o maior grau de separação (desvio padrão) possível (FISHER,1936).

Desta forma, a FDLF procura por uma combinação linear das características observadas que proporcione maior poder de discriminação entre os grupos, tendo como propriedade minimizar as probabilidades de má classificação.

Silva *et al.* (2012) avaliaram oito características da qualidade das mudas e usaram os dados transformados através da FDLF. Os autores concluíram com o estudo das características que há diferenças na influência da composição dos substratos nas cultivares e com a análise da FDLF que o manejo convencional é superior ao orgânico.

A combinação linear do vetor *x*, $Y = \hat{a}x$, em cada população, de maneira que seja o máximo da relação do quadrado da diferença de médias dos conjuntos *A* e *B* (x_A e x_B) com à sua variância *Y*, ou seja, que fornece o máximo para a proporção. Neste contexto, a FDLF amostral, é dada a seguir pela equação (3.10).

$$Y = (x_A - x_B)' S_p^{-1} x \tag{3.10}$$

em que *x* = vetor das variáveis aleatórias correspondentes às características amostrais observadas.

- Se $x_0 \in A$, então:

  $y_0 = (x_A - x_B)' S_p^{-1} x_0 \geq q = \frac{1}{2}(x_A - x_B)' S_p^{-1}(x_A + x_B)$

- Se $x_0 \in B$, então:

  $y_0 = (x_A - x_B)' S_p^{-1} x_0 < q = \frac{1}{2}(x_A - x_B)' S_p^{-1}(x_A + x_B)$

3.1.4.4 Redes Neurais Artificiais (RNAs)

A Rede Neural Artificial (RNA) é formada por um conjunto de neurônios que interagem entre si, similar ao funcionamento dos neurônios do cérebro humano. Basicamente, são sistemas computacionais com processamento altamente paralelo e distribuído e que apresentam a habilidade de aprender e armazenar através de um conjunto de dados, este método soluciona problemas através da simulação do cérebro humano, inclusive em seu comportamento (GUO; HU; YI, 2004). As RNAs têm sido utilizadas em sistemas de controle e otimização, análise de aplicações financeiras, reconhecimento de voz, classificação, RP e entre outros.

As RNAs são compostas de muitos elementos simples, inspirados pelo sistema nervoso biológico, que operam em paralelo. A função da rede é determinada pelas conexões entre os seus elementos. Pode-se treinar uma rede neural para executar uma função particular ajustando-se os valores das conexões entre os elementos (STEINER, 2014).

A RNA pode ser vista como um conjunto de unidades de entrada e saída conectadas por camadas intermediárias e cada ligação possui um peso associado onde realizam diversos processamentos, conforme ilustrado na Figura 3.5.

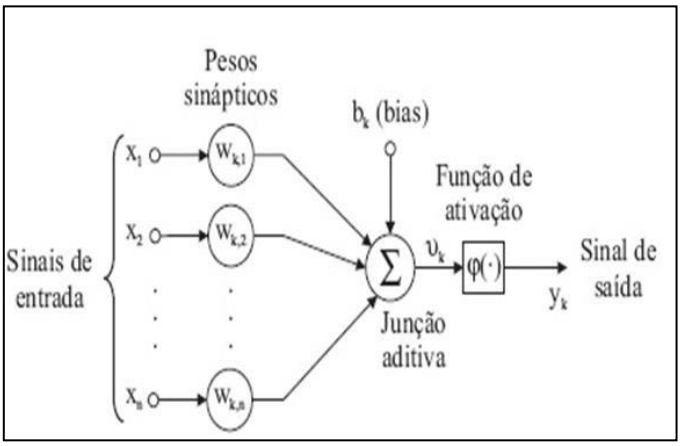

Figura 3.5 - Modelo de neurônio artificial
Fonte: Haykin (2009)

Segundo Haykin (2009) podem-se identificar três elementos básicos na estrutura de um neurônio artificial:

- As sinapses que recebem os sinais de entrada;
- O somatório para somar os sinais de entrada com seus respectivos pesos sinápticos;
- A função de ativação, que restringe a amplitude da saída do neurônio.

As entradas são conectadas aos elementos processadores básicos, que são por sua vez interconectados com elementos de outras camadas e/ou a saída da rede. O número de neurônios na camada escondida é definido de acordo com o número de vetores utilizados para o treinamento. Na camada de saída, a quantidade de neurônios dependerá, em geral, do número de classes existentes.

Segundo Haykin (2009) o neurônio pode ser descrito matematicamente, pelas seguintes equações (3.11) e (3.12):

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k \qquad (3.11)$$

$$y_k = \varphi(v_k) \tag{3.12}$$

onde:

x_1, x_2, ... , x_m são os sinais de entrada; w_{k1}, w_{k2}, ... , w_{km} são os pesos sinápticos; v_k é o campo local induzido; b_k é o *bias*; $\varphi(.)$ é a função de ativação; e y_k é o sinal de saída, todos relacionados ao neurônio k. Além disso, o parâmetro *bias* ainda pode ser representado pelo peso sináptico de uma entrada cujo valor é fixo em “1”.

- Tipos de função de ativação

A função de ativação é muito importante para o comportamento de uma RN porque é ela que define a saída do neurônio artificial e, portanto, o caminho pelo qual a informação é conduzida. A Figura 3.6 a seguir apresenta exemplos de funções de ativação.

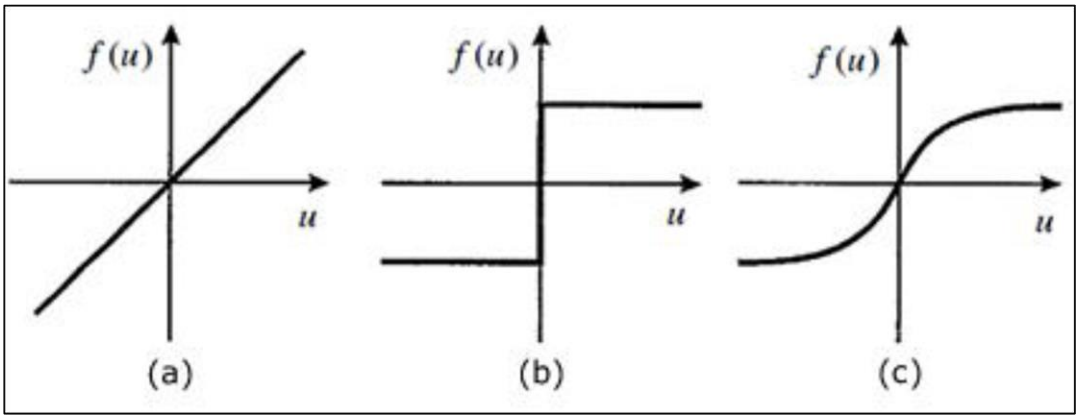

Figura 3.6 - Funções de ativação: (a) função linear, (b) função *threshold*, (c) função sigmoidal
Fonte: Haykin (2009)

Estes três principais tipos de função de ativação, $\varphi(.)$, que restringe a saída do neurônio pode ser expressa matematicamente pelas seguintes funções (3.13; 3.14; 3.15):

- Função Linear:

$$\varphi(v) = v \tag{3.13}$$

- Função Threshold:

$$\varphi(v) = \begin{cases} 1, se\ v \geq 0 \\ 0, se\ v < 0 \end{cases} \tag{3.14}$$

- Função Sigmoidal:

$$(v) = \frac{1}{1+\exp(-av)} \tag{3.15}$$

- Tipos de treinamento de uma RN

A RNA é capaz de “aprender sozinha”, por um período de treinamento apropriado e podem generalizar os resultados obtidos para dados previamente desconhecidos, ou seja, produzir respostas coerentes e apropriadas para atributos ou exemplos que não foram utilizados no seu treinamento (ZHANG, 2010).

As RNAs possuem a capacidade de generalização a partir da apresentação de um conjunto de amostras contendo entradas e suas respectivas saídas. O treinamento em uma RN, pode ser, em geral, de dois tipos:

- **Supervisionado:** quando a saída é dada para um conjunto de entradas e o sucesso é obtido quando se obtém a correta saída para a correspondente entrada. Em um treinamento supervisionado cada vetor de entrada é associado com um correspondente vetor de saída. A rede é treinada utilizando estes pares de entradas e saídas. O processo de treinamento ou aprendizado supervisionado de uma rede neural consiste, essencialmente, em minimizar o erro entre a saída da rede para um determinado padrão de entrada e a resposta esperada para aquele mesmo padrão (RODRIGUES, 2006; STEINER, 2014).
- **Não-Supervisionado:** quando somente um conjunto de entradas é dado e tem-se que classificá-los, extraindo quaisquer propriedades estatísticas, de acordo com algumas representações internas. O algoritmo de treinamento para a rede não-supervisionada, desenvolvido por Kohonen, em 1984, modifica a configuração da rede para agrupar vetores de entrada similares em classes extraindo propriedades estatísticas do conjunto de treinamento. Outro modelo baseado na filosofia de aprendizagem não-supervisionada é a Rede de Hopfield. Redes treinadas com algoritmos de aprendizado não supervisionado são empregadas em problemas de agrupamento (*clustering*) e mineração (*mining*) de dados (RODRIGUES, 2006; STEINER, 2014).

- Fluxo de Dados em uma RN

Segundo Steiner (2014) “em função do fluxo de dados as redes neurais podem ser classificadas em: redes *feed-forward*, se elas podem propagar os dados apenas unidirecionalmente, ou seja, apenas para a frente; ou redes *feedback* ou recorrentes, se o fluxo de dados pode se dar nos dois sentidos”.

- Aprendizado em uma RNA

Em uma RNA, o aprendizado ocorre à medida que os pesos sinápticos são ajustados com base em alguma regra pré-estabelecida, como a regra delta (ou regra de Widrow-Hoff).

Desta maneira, durante o processo de aprendizagem os pesos normalmente “sofrem” uma modificação iterativa. No decorrer das iterações, os pesos influenciam os cálculos realizados pela rede. O algoritmo de aprendizagem julga a qualidade do peso de acordo com os valores obtidos pela saída da rede. Matematicamente, podemos definir a atualização dos pesos da rede na iteração n pela seguinte equação (3.16) (RAUBER, 2011).

$$w_{kj}(n+1) = w_{kj}(n) + \Delta w_{kj}(n) \tag{3.16}$$

onde:

w_{kj} e Δw_{kj} representam os pesos da rede e suas variações.

Os algoritmos de aprendizado diferem na forma como Δw_{kj} é calculado (BRAGA; CARVALHO; LUDERMIR 2011). Segundo a regra delta, o ajuste Δw_{kj} aplicado ao peso sináptico w_{kj} é definido por (3.17) (HAYKIN, 2001).

$$\Delta w_{kj}(n) = \eta e_k(n) x_j(n) \tag{3.17}$$

onde:

η = taxa de aprendizado; e_k = sinal de erro, ou seja, a diferença entre o valor esperado; x_j = sinal de entrada. Além disso, pode ser definido o cálculo do erro, conforme equação (3.18).

$$e_k(n) = d_j(n) - y_{kj}(n) \tag{3.18}$$

onde:

d_j = sinal efetivo de saída da rede y_k.

O sinal de erro aciona um mecanismo de controle cujo propósito é aplicar uma sequência de ajustes corretivos aos pesos sinápticos do neurônio *k*. Os ajustes corretivos dos pesos são projetados para aproximar passo a passo o sinal de saída y_k da resposta desejada d_j". Para tanto, usa-se minimizar a função do erro quadrático médio, definida em termos do sinal do erro segundo a equação (3.19) (BRAGA; CARVALHO; LUDERMIR, 2011; HAYKIN, 2001).

$$\varepsilon(n) = \frac{1}{2} \sum_{k \in C} e^2_k (n) \tag{3.19}$$

onde:

o conjunto C inclui todos os neurônios da camada de saída da rede. A partir de ε, o conjunto de dados formado pelos pares de entrada e saída $[x_j;\ d_j]$ define a superfície de erro.

- Modelos de Redes Neurais

São muitos os modelos de RNAs, sendo que aqui são apresentados os modelos básicos: o Perceptron, Redes Lineares e Redes de Múltiplas Camadas.

- **Perceptron**

Segundo Haykin (2009), "o *perceptron* de camada única é a forma mais simples de uma rede neural usada para a classificação de padrões linearmente separáveis".

Para Steiner (2014) "a rede Perceptron consiste de uma única camada de *i* neurônios conectados as η entradas através de um conjunto de pesos *w*(*i*, *j*). Os índices da rede, *i* e *j*, indicam que *w*(*i*, *j*) é a "força" da conexão da *j-ésima* entrada ao *i-ésimo* neurônio".

De um modo resumido, temos o algoritmo do *Perceptron* de camada única, com k neurônios (BRAGA; CARVALHO; LUDERMIR, 2011; HAYKIN, 2009; STEINER, 2014):

- ✓ Inicialização: faça o vetor de pesos sinápticos $w_k(n) = 0$ e o bias $b_k(n) = 0$.

- ✓ Ativação: execute o somatório $u_k(n) = \sum_{j=1}^{m} w_{kj}(n)x_j(n) + b_k$, onde m é a quantidade de pesos sinápticos para o neurônio k; execute a função de transferência $y_k = \varphi(u_k(n))$, onde $\varphi(.\ )$ é uma função de limiar, linear ou sigmoide, entre outras; calcule o sinal de erro $e_k(n) = d_j(n) - y_k(n)$, onde d_j é o valor esperado.
- ✓ Atualização dos pesos: calcule o novo peso sináptico $w_{kj}(n+1) = w_{kj}(n) + \Delta w_{kj}(n)$, onde $\Delta w_{kj}(n) = \eta e_k(n)x_j(n)$, com $0 < \eta < 1$, $x_j(n)$ e $0 < j < m$ ($w_{k0} = b_k$ e $x_0 = 1$) correspondendo aos sinais de entrada.
- ✓ Análise da situação atual: após todos os exemplos de treinamento terem passado pela rede uma única vez teremos findado a 1ª iteração, então calcule o erro global $\varepsilon(n) = \frac{1}{2}\sum_{k \in C} e_k^2(n)$, para o conjunto C dos neurônios de saída da rede.
- ✓ Continuação: incremente a iteração n em uma unidade e volte ao processo de ativação.

O procedimento deverá continuar até que um erro ε suficientemente pequeno seja encontrado.

- **Redes Lineares**

Segundo Steniner (2014) “estas redes diferem do Perceptron na função de tranferência que é linear, permitindo que as saídas tomem qualquer valor entre "0" e "1" e não apenas os valores "0" e "1" como no Perceptron”.

Estas redes utilizam a regra de aprendizagem de Widrow-Hoff, também conhecida como a regra dos Mínimos Quadrados, ajusta os pesos das conexões entre os neurônios da rede de acordo com o erro, ou seja, esta regra tem como objetivo encontrar um conjunto de pesos e polarizações que minimizem a função erro (BATISTA, 2012). A regra de Widrow-Hoff pode somente treinar redes lineares de uma única camada. Esta rede é conhecida como *Madaline* (*Adaline* de múltiplas camadas) por conter muitas *Adalines*.

A rede *Madaline* ou *Adaline* tem uma camada de entrada com N unidades e uma camada de saída com apenas uma unidade. Não há camadas escondidas. O número de entradas da rede e o número de neurônios na camada de saída estão restritos pelo número de entradas e saídas exigidos pelo problema. A atividade do neurônio de saída não é uma variável binária como no Perceptron, mas uma função

linear do seu nível de ativação (HAYKIN, 2009; LIMA; PINHEIRO; SANTOS, 2014; STEINER, 2014). Na Figura 3.7 tem-se a representação esquemática de uma Rede Linear (observar que a função de transferência é linear).

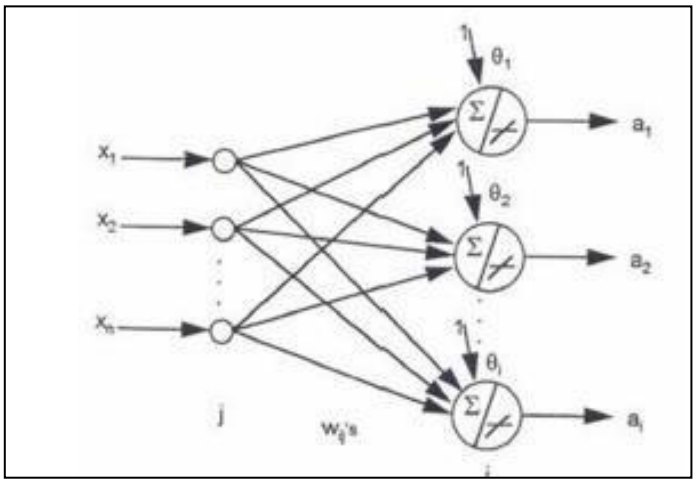

Figura 3.7 - Rede Linear
Fonte: Steiner (2014)

- **Perceptron de múltiplas camadas**

Redes neurais multicamadas são arquiteturas em que neurônios são organizados em duas ou mais camadas, nada mais é que uma generalização do *Perceptron* de camada única (HAYKIN, 2009). Podemos descrevê-lo como uma rede constituída pela camada de entrada, uma ou mais camadas ocultas e uma camada de saída; rede essa que apresenta um alto grau de conectividade, a Figura 3.8 representa um esquema típico de uma rede neural artificial com múltiplas camadas (BRAGA; CARVALHO; LUDERMIR, 2011; HAYKIN, 2009).

Seu treinamento se dá em três etapas: a alimentação para frente (*feedforward*) da rede com os padrões de entrada, o cálculo e retropropagação (*backpropagation*) do erro, e o ajuste dos pesos (FAUSETT, 1994).

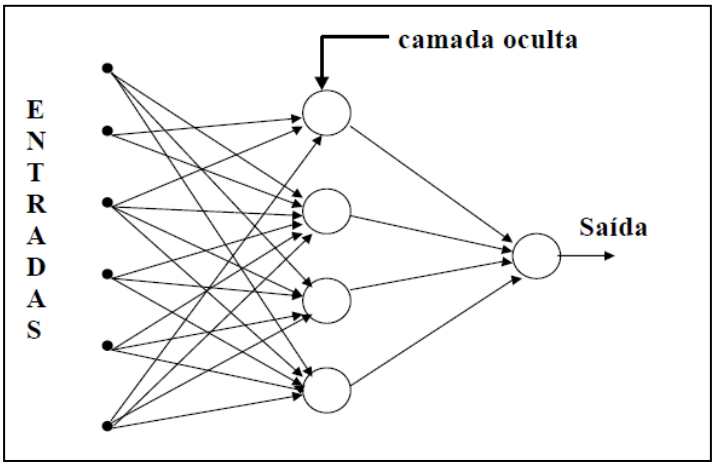

Figura 3.8 - Redes Multicamadas
Fonte: Haykin (2009)

O termo *backpropagation* surge do fato que o algoritmo se baseia na retropropagação dos erros para realizar os ajustes de pesos das camadas intermediárias. A forma de calcular as derivadas parciais do erro de saída em relação a cada um dos pesos da rede é o que caracteriza o *backpropagation* (REZENDE,2003).

O *backpropagation* é considerado um algoritmo de treinamento do tipo supervisionado, e é o mais popular para o treinamento de *perceptrons* de múltiplas camadas, já que utiliza as informações dos padrões de entrada fornecidos à rede e a sua respectiva saída desejada e que, através de um mecanismo de correção de erros (gradiente descendente), ajusta os pesos da rede aos padrões de entrada fornecidos na etapa de treinamento (BRAGA; CARVALHO; LUDERMIR, 2011).

Assim sendo, o treinamento é realizado em duas etapas: (i) a fase *forward* e (ii) a fase *backward*. Na 1ª fase, são apresentados conjuntos de dados à rede e está propaga o sinal até a sua camada de saída. A 2ª fase realiza a correção dos erros, alterando os pesos internos da rede, conforme a saída desejada (BRAGA; CARVALHO; LUDERMIR, 2011).

3.1.4.5 Máquina de Vetor Suporte / *Support Vector Machine* (SVM)

A SVM é uma técnica de aprendizagem supervisionada que analisa e reconhece dados utilizados em problemas de classificação e de regressão. Esta técnica é treinada com um algoritmo baseado na teoria estatística de aprendizagem, cujos vetores do espaço de entrada são mapeados não linearmente em um espaço

com características de alta dimensionalidade, através de um mapeamento escolhido a priori (função *kernel*).

Neste espaço de características, é construída uma superfície de decisão linear, que se constitui de um hiperplano de separação ótimo e que apresenta propriedades especiais que garantem alta habilidade de generalização da máquina de aprendizagem (CRISTIANINI; SHAWE-TAYLOR, 2002).

Segundo Haykin (2009), a equação do hiperplano que separa o conjunto de dados é apresentada a seguir pela equação (3.20).

$$w^T x + b = 0 \tag{3.20}$$

onde:

x = vetor de entrada; w = vetor de pesos e $b \in R$.

Assim sendo, podem ser definidas as duas regiões que contêm cada uma das classes, através da equação (3.21).

$$f(x) = \begin{cases} se\ w^T + b \geq então + 1 \\ se\ w^T + b < então - 1 \end{cases} \tag{3.21}$$

onde:

$f(x)$ mapeia cada um dos dados (exemplos) em uma das classes {+1, -1}.

Seja x_1 um ponto no hiperplano H_1: $w \,.\, x + b = +1$ e x_2 um ponto no hiperplano H_2: $w \,.\, x + b = -1$, conforme ilustrado na Figura 3.9. Projetando $x_1 - x_2$ na direção de w, perpendicular ao hiperplano separador $w \,.\, x + b = 0$, é possível obter a distância entre os hiperplanos H_1 e H_2, através da seguinte equação (3.22) (CAMPBELL, 2000).

$$d = (x_1 - x_2)\left(\frac{w}{||w||} \cdot \frac{(x_1 - x_2)}{||x_1 - x_2||}\right) \tag{3.22}$$

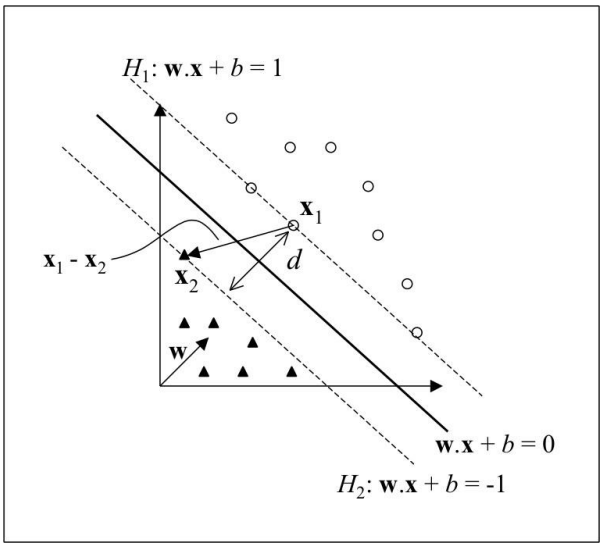

Figura 3.9 - Distância *d* entre os hiperplanos H_1 e H_2
Fonte: Lorena e Carvalho (2007)

Tem-se que $w \,.\, x_1 + b = +1$ e $w \,.\, x_2 + b = -1$. A diferença entre essas equações fornece $w \,.\, (x_1 - x_2) = 2$ (SCHÖLKOPF e SMOLA, 2002). Substituindo esse resultado na equação (3.22), tem-se a equação (3.23).

$$d = \frac{2\,(x_1 - x_2)}{||\mathrm{w}||\;||x_1 - x_2||} \tag{3.23}$$

Como se deseja obter o comprimento do vetor projetado toma-se a norma da equação (3.23), obtendo-se a equação (3.24).

$$d = \frac{2}{||\mathrm{w}||} \tag{3.24}$$

Esta é a distância *d* ilustrada pela a Figura 3.9, entre os hiperplanos H_1 e H_2, paralelos ao hiperplano separador. Como *w* e *b* foram escalados de forma a não haver exemplos entre H_1 e H_2, $1/||w||$ é a distância mínima entre o hiperplano separador e os dados de treinamento.

O problema de identificação do hiperplano de separação das classes pode ser definido como um problema de otimização como em (3.25) e (3.26). Verifica-se que a maximização da margem de separação dos dados em relação a $w \,.\, x + b = 0$, pode ser obtida pela minimização de $||w||$ (SCHÖLKOPF; SMOLA, 2002; VIEIRA, 2006; LORENA; CARVALHO, 2007).

$$Minimizar = \frac{1}{2}||w||^2 \tag{3.25}$$

$$\text{s.a: } y_i\,(w\,.\,x_i\, + b) - 1\, \geq\, 0\,,\;\; i\, =\, 1, 2, \ldots, n \tag{3.26}$$

Devido à natureza das restrições do problema primal, pode-se ter dificuldades na obtenção da solução do problema. Conforme ALES *et al.* (2009), os problemas deste tipo podem ser melhor trabalhados utilizando-se do método de *Lagrange*, em sua forma dual, e assim obtendo o seguinte problema de otimização, conforme o modelo matemático (3.27) a (3.29).

$$Maximizar\ \alpha = \Sigma_{i=1}^{n}\ \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\ \alpha_i\ \alpha_j\ y_i\ y_j\ (x_i\ .\ x_j) \tag{3.27}$$

$$\text{s.a: } \alpha_i \geq\, 0\,,\;\; i\, =\, 1, 2, \ldots, n \tag{3.28}$$

$$\Sigma_{i=1}^{n}\ \alpha_i\, y_i = 0 \tag{3.29}$$

onde:

α é o vetor que representa os multiplicadores de *Lagrange* e *α** é a solução ótima para o problema dual.

A solução do problema primal é dada por *w** e *b**. Conhecendo-se o valor de *α**, pode-se calcular o valor de *w**, conforme a equação (3.30) (LORENA; CARVALHO, 2007):

$$w = \Sigma_{i=1}^{n}\ \alpha_i y_i x_i \tag{3.30}$$

Essa formulação é denominada forma dual, enquanto o problema original é referenciado como forma primal. Também é importante notar que no ponto de sela, para cada Multiplicador de *Lagrange* α_1, o produto daquele multiplicador pela restrição correspondente desaparece. O cálculo de *b** pode ser realizado utilizando o valor de *α** e as condições de *Karush-Kuhn-Tucker*, e consequentemente temos a equação (3.31) (LORENA; CARVALHO, 2007):

$$\alpha_i^*[y_i(w^*\,.\ x_i\ b^*) - 1] = 0\,,\;\; i\, =\, 1, 2, \ldots, n \tag{3.31}$$

Denominam-se Vetores Suporte aqueles vetores que se encontram justamente nos hiperplanos de separação dos dados, de tal modo, $\alpha^*_i > 0$. Pontos dos vetores suporte não participam do cálculo de *w**. Eliminando as variáveis primais do problema de otimização, a equação que define o hiperplano de separação é dada pela seguinte equação (3.32) (VIEIRA, 2006):

$$f(x) = sinal\ (\ \Sigma_{i=1}^{n}\ y_i\alpha_i\ (x_i + b)) \tag{3.32}$$

onde:

b é calculado utilizando um vetor suporte e a equação (3.31).

Em situações reais, a teoria sobre SVM descrita até este ponto não é suficiente para se obter bons resultados, ou seja, é muito comum haver interseções entre as classes e ainda podem ocorrer ruídos (*outliers*) nos dados ou da própria natureza do problema, que pode ser não linear. Para realizar essa tarefa, permite-se que alguns dados possam violar as restrições da equação (3.26) e isto é feito com a inclusão de variáveis de folga na superfície de separação dos dados, assim a equação (3.25) e (3.26) pode ser reescrita da seguinte maneira (IVANCIUC, 2007; HAYKIN, 2009).

$$Minimizar = \frac{1}{2}||w||^2 + C\ (\ \Sigma_{i=1}^{n}\ \xi_i) \tag{3.33}$$

$$s.a:\ yi\ (w\ .xi\ +\ b)\ \geq\ 1\ \xi_i, \xi_i\ \geq\ 0,\ \ i\ =\ 1, 2, \ldots, n \tag{3.34}$$

onde:

é a variável de folga e C é um parâmetro que visa ponderar a minimização do erro no conjunto de treinamento. Um bom valor de C precisa ser identificado a partir de testes com dados reais, e varia de acordo com a natureza do problema (HAYKIN, 2009).

O problema dual possui a mesma solução do primal, ou seja, pode-se revolver o problema primal indiretamente através da resolução do problema dual. O Lagrangeano é obtido por meio de Multiplicadores de Lagrange, que estão associados às restrições de desigualdade do problema primal. Esses multiplicadores de Lagrange devem ser positivos. As restrições do problema dual ficam mais simples do que as do problema primal, sendo dado pela seguinte equação (3.35) a (3.37).

$$Max\ L\ (w, b, \xi, \alpha) = \Sigma^{n}_{i=1}\ \alpha_i - \frac{1}{2}\ \Sigma^{n}_{i=1}\ \Sigma^{n}_{j=1}\ \alpha_i\ \alpha_j\ y_i\ y_j\ (x^i)(x^j)) \tag{3.35}$$

$$s.a: \Sigma^{n}_{i=1}\ \alpha_i\ y_i = 0 \tag{3.36}$$

$$0 \leq \alpha_i \leq C \tag{3.37}$$

Conforme já mencionado, em alguns problemas reais os padrões a serem classificados podem não ser linearmente separáveis. Na formulação do problema dual existe um produto interno $(x^i)(x^j)$ que pode ser substituído por uma função $K(x^i,\ x^j)$, denominada função *kernel*, obedecendo o Teorema de Mercer (SCHÖLKOPF; SMOLA, 2002; CRISTIANINI; SHAWE-TAYLOR, 2002). Essas funções *kernel* mapeiam os dados de entrada em um espaço de dimensão maior, no qual os dados podem ser separados através de um hiperplano, ou seja, os dados tornam-se linearmente separáveis.

A Figura 3.8, como pode ser observado, apresenta um exemplo cujos padrões estão no seu espaço original, onde não é possível criar uma fronteira linear de separação das classes. Em (a) estão todos os padrões identificados por classe e em (b) é mostrada a fronteira não linear de separação dos mesmos. Pode-se utilizar funções não lineares nos vetores de entrada a fim de mapeá-los em um espaço intermediário com mais dimensões, de maneira tal que seja possível identificar um hiperplano linear ótimo que separe as classes e em (c) mostra o resultado desta transformação, além do hiperplano (VIEIRA, 2006).

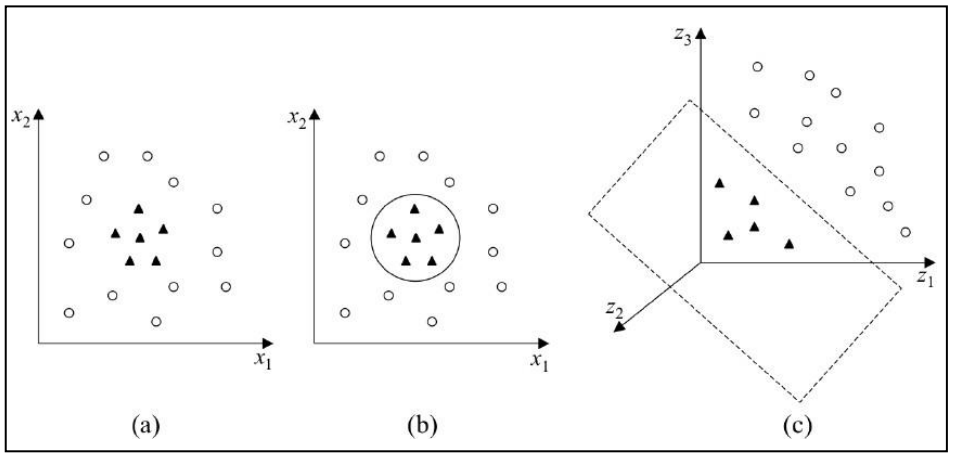

Figura 3.10 - Conjunto de dados não linearmente separáveis
Fonte: Lorena e Carvalho (2007)

Assim sendo, a transformação do vetor de entrada pode ser dada por meio do núcleo (*kernel*) do produto interno deste vetor com um vetor suporte. Este núcleo pode ser polinomial, RBF (*Radial-Basis Function*/Gaussiana de Funções Radias) ou ainda *Perceptron* de duas camadas (VEIRA, 2006; HAYKIN, 2009).

As funções *kernel* realizam um mapeamento implícito, portanto nas formulações do SVM os exemplos de treinamento nunca aparecem isolados, mas em forma de produto interno, que pode ser substituído por uma função *kernel*. As funções *Kernel* mais utilizadas são dadas na tabela 3.1.

Quadro 3.1 - Conjunto de dados não linearmente separáveis

Função de *Kernel*	Expressão para k (x_i, x_j)	Parâmetros
RBF	$e^{\|\|x_i - x_j\|\|^2/2\sigma^2}$	σ^2
Polinomial	$(x_i^T x_j + a)^2$	a, b
Sigmóide	$\tanh(\beta_0\, x_i^T x_j + \beta_1)$	β_0, β_1

Assim sendo, os parâmetros desta função e do algoritmo de determinação do hiperplano ótimo e a escolha do *Kernel* e de seus parâmetros afetam o desempenho do classificador através da superfície de decisão.

Além disso, a finalidade da SVM é separar as instâncias por meio da etapa de treinamento com o objetivo de produzir um classificador que funcione de maneira adequada com exemplos não identificados (etapa de teste), ou seja, exemplos que não foram aplicados durante o treinamento, adquirindo assim a capacidade de predizer novos exemplos.

3.2 TRABALHOS CORRELATOS

Esta seção relata alguns trabalhos relacionados com o tema desta pesquisa e traz um resumo de como autores têm trabalhado com as técnicas de KDD aqui apresentadas em diferentes problemáticas e seus resultados.

Steiner e Carnieri (1999) utilizaram dados de aproximadamente 10.000 clientes obtidos a partir de um grande banco privado brasileiro, e apresentaram uma metodologia para realizar a análise de *Credit Scoring*. A metodologia proposta foi dividida em duas fases: análise estatística dos dados (Correlação entre os dados e T^2 de Hotelling) e a utilização de um modelo (RL) para realizar o RP. A técnica

utilizada para RL foi a GLIM (*Generalized Linear Interactive Model*/Modelo Linear Generalizado Interativo). Os dados foram modelados para clientes com renda igual ou inferior a R$: 1.000,00. O desempenho do modelo foi considerado razoável, apresentando cerca de 20% de erros. Esta percentagem relativamente elevada pode ser explicada pela falta de informações comportamentais de cada cliente e pela grande variabilidade existente na aplicação de informações de dados.

Lemos, Steiner e Nievola (2005) analisaram registros históricos de clientes (pessoas jurídicas) de uma agência bancária, por meio de duas ferramentas a Redes Neurais e a Árvores de Decisão. Essas técnicas permitiram fazer o RP e também classificar novos atributos. Na implementação da técnica Árvores de Decisão optou-se por utilizar o *software* computacional *Weka* (*Waikato Environment for Knowledge Analysis*/Waikato Ambiente para a Análise do Conhecimento) e para a implementação da técnica de Redes Neurais ao problema, foi utilizado o *software MATLAB* (*Neural Networks Toolbox*). Os resultados foram bastante satisfatórios, mostrando que, para esse problema específico, as Redes Neurais apresentaram uma taxa de classificação correta maior do que aquela das Árvores de Decisão.

Steiner *et al*. (2006) mostraram a influência da análise exploratória dos dados no desempenho das técnicas DM quanto à classificação de novos atributos por meio da sua aplicação a um problema médico, além de comparar o desempenho delas entre si, visando obter a técnica com o maior percentual de acertos. As técnicas utilizadas foram: a GSME, a Programação Linear (PL), a FDLF, a RL, a Redes Neurais e a Árvores de Decisão. Para a análise exploratória de dados foram: o Teste T^2 de Hotelling, a transformação dos atributos e o descarte de pontos atípicos.

No entanto, a técnica que envolve um modelo de PL e uma outra que envolve Redes Neurais foram as técnicas que apresentaram os menores percentuais de erros para os conjuntos de testes, apresentando capacidades de generalização satisfatórias. Assim, pode-se concluir que os métodos apresentados podem ser utilizados nos mais diversos problemas reais de classificação e que a referida análise, se conduzida de forma adequada, pode trazer importantes melhorias nos desempenhos de quase todas as técnicas abordadas, tornando-se, assim, uma importante ferramenta para a otimização dos resultados finais.

Steiner *et al*. (2007) apresentaram ferramentas que podem ajudar a identificar e prever quais clientes serão bons pagadores (ou não) de crédito junto a bancos. Um conjunto de dados de crédito foi analisado usando a técnica de extração de

regras *NeuroRule*, a partir de uma RNA treinada. Para tanto, os autores fizeram uso do *software* Weka (*Waikato Environment for Knowledge Analysis*). Os resultados foram considerados bastante satisfatórios, alcançando mais de 80% de acurácia, destaque foi dado ao fato de que os resultados por meio de regras ficam mais claros aos especialistas (gerentes de crédito).

A técnica de RNA foi também utilizada por Baptistella, Cunico e Steiner (2007) para a estimação dos valores venais de imóveis urbanos para o município de Guarapuava, PR. Para tanto, utilizaram-se de dados do Cadastro Imobiliário fornecidos pelo setor de Planejamento da Prefeitura Municipal. A técnica da Análise das Componentes Principais (ACP) foi usada para reduzir e transformar as variáveis originais em nove fatores. Assim, os resultados da amostra de dados completa foram comparados com os resultados obtidos dividindo-se a mesma em grupos menores, compostos de bairros com características semelhantes, sendo que a acurácia com estes últimos (grupos menores) foram superiores aos obtidos com o primeiro (amostra completa).

Steiner *et al*. (2008) apresentaram uma metodologia, composta por técnicas de Análise Multivariada, para a construção de um modelo estatístico de Regressão Linear Múltipla (RLM) para avaliação de imóveis em função de suas características (variáveis, atributos). Foi aplicada a análise de agrupamento para obtenção de grupos homogêneos dentro de cada classe e a técnica da ACP para resolver o problema da multicolinearidade que pode existir entre as variáveis do modelo. O modelo para cada grupo homogêneo, dentro de cada classe de imóveis avaliados, apresentou um ajuste adequado aos dados e uma capacidade preditiva bastante satisfatória.

Uma proposta para controle de qualidade de bobinas de papel em uma indústria de papel foi apresentada por Steiner, Carnieri e Stange (2009). Através da obtenção de dados quantitativos representativos das características que compõem bobinas de papel de boa e baixa qualidade, utilizaram as seguintes técnicas: Geração de uma Superfície Linear por Partes (GSLP-PL), GSME-PL, RLB, FDLF e método dos *k*-Vizinhos Mais Próximos (K-NN) segundo a distância de *Mahalanobis*. Na análise exploratória dos dados, os autores utilizaram o teste T^2 de Hotelling. Para os métodos de PL utilizaram a linguagem Pascal através o pacote computacional GAMS (*General Algebraic Modeling Systems*/Sistema Geral de *Modelagem Algébrica*), e o pacote computacional GLIM para a RL e na análise exploratória dos

dados. Dando prosseguimento, os autores utilizaram o GSME-PL (apresentou a maior acurácia), para a construção de um modelo matemático que permitisse ajustar as variáveis no decorrer do processo de fabricação do papel industrial, de modo a garantir que as bobinas de papel sejam sempre de boa qualidade e a um mínimo custo.

Já Bettiollo e Steiner (2009) compararam o desempenho das técnicas de RNA e de RLB na classificação de instâncias apresentadas pela Avaliação Bioquímica de Risco Fetal (ABRF) com relação a síndrome de *Down*. As duas técnicas foram capazes de, após serem treinadas e testadas, identificar o nível do risco (alto ou baixo) do feto ser portador da doença.

Os resultados para as RNAs, assim como para a RL, foram obtidos através do *software Statgraphics*. Estas duas técnicas de RP foram bastante eficientes na tarefa de classificação dos atributos apresentados, sendo que as RNAs classificaram corretamente cerca de 93% das instâncias do conjunto de treinamento e cerca de 85% das instâncias do conjunto de teste. Estes percentuais para a técnica de RL foram de 93% e 86%, respectivamente.

Martínez-López e Casillas (2009) propuseram uma nova e completa metodologia para o KDD, a ser aplicada na modelagem de *marketing* casual para ser usada como uma ferramenta de apoio à decisão de gestão de *marketing*. A metodologia é baseada em Sistemas *Fuzzy* Genéticos, uma hibridização específica de métodos de Inteligência Artificial, altamente adequado para o problema da pesquisa abordado pelos autores. Este sistema inteligente permite ao pesquisador obter uma visão das relações entre as variáveis de uma nova forma, quando comparada com o tipo de saída que os autores usaram anteriormente para obter as relações a partir de técnicas estatísticas.

A metodologia foi aplicada a um problema real de modelagem do comportamento do consumidor em ambientes *on-line*, onde os autores puderam oferecer uma perspectiva geral de como ele funciona. Os resultados que obtiveram foram satisfatórios. Além disso, os atributos são mostrados de forma a facilitar a compreensão, além de permitir encontrar cenários interessantes nos bancos de dados os clientes analisados.

Por ocasião da publicação do artigo, o *software* desenvolvido ainda não estava pronto para ser comercializado e segundo os autores ainda seriam

necessárias melhorias no algoritmo genético (AG) para melhorar a precisão dos resultados e de desempenho.

Segundo Fang e Rachamadugu (2009), o KDD fornece às organizações ferramentas necessárias para filtrar grandes armazenamentos de dados para extrair conhecimento. Este processo suporta e melhora a tomada de decisão nas organizações. Consequentemente, apresentaram e definiram o conceito de atualizar o conhecimento, um passo fundamental para garantir a qualidade e a pontualidade do conhecimento descoberto no processo KDD. Em consequência, estudaram o atualizador do conhecimento partindo da perspectiva de que quando atualizar o conhecimento o custo total do sistema ao longo de um espaço de tempo seja minimizado.

Os autores propuseram um modelo para atualizar o conhecimento e uma metodologia de Programação Dinâmica para o desenvolvimento de estratégias ótimas. A metodologia de Programação Dinâmica proposta tem baixo custo computacional, e também tem baixa exigência de "*state space*" (um conjunto de valores que um processo pode assumir), podendo ser facilmente implementada, na prática, com base em parâmetros tais como a receita vinda do conhecimento (dinheiro gerado pelo conhecimento). Os autores demonstraram a eficácia da metodologia proposta utilizando dados de uma aplicação do mundo real. A metodologia proposta fornece aos tomadores de decisão uma orientação na execução do KDD de forma eficaz e eficiente.

Mendes, Fiuza e Steiner (2010) analisaram a relevância dos dados coletados através de questionários respondidos por 2177 pacientes com diagnóstico de dor de cabeça e, através desta análise, verificar se o tratamento desses dados usando RNAs como ferramenta de RP pode auxiliar nos diagnósticos de novos pacientes.

O sistema desenvolvido é baseado em RNAs do tipo *Perceptron* multicamadas e utilizou-se o software *MATLAB* 7.0 e o componente *Neural Network Toolbox* para a sua implementação e seu treinamento.

Desta forma, foram levantados elementos para justificar a utilização de RNAs como ferramenta de apoio ao diagnóstico, objetivando auxiliar o médico no seu dia-a-dia, e também como uma ferramenta educacional de auxílio ao treinamento e qualificação de profissionais da área médica. Os resultados obtidos foram bastante satisfatórios, mostrando que as RNAs podem ser eficientes na resolução deste problema específico.

Guruler, Istanbullu e Karahasan (2010) apresentaram uma descoberta de conhecimento aplicado sobre os dados demográficos dos estudantes universitários. A fim de explorar os fatores que têm impacto sobre o sucesso de estudantes universitários, um *software* de descoberta de conhecimento, chamado MUSKUP (*Mugla University Student Knowledge Discovery Unit Program*), foi desenvolvido e testado em dados dos alunos. Neste sistema a classificação de Árvore de Decisão é utilizada como uma técnica de DM.

Com este sistema, todas as tarefas envolvidas no processo de descoberta de conhecimento são realizadas coletivamente. A vantagem dessa abordagem é ter acesso a todas as funcionalidades do SQL *Server* e *Analysis Services* através de um *software* único.

Ao avaliar o desempenho dos alunos, a técnica de classificação de Árvore de Decisão foi realizada. As classificações mostram quais os dados demográficos que influenciam no GPA (*Grade Point Averages*/Média de Notas) dos estudantes. O escopo do estudo se limita a determinar os perfis de estudantes cuja GPA é igual a 2.0 (que é a média mínima exigida para a graduação) ou, igual a 3,0 ou superior (diploma de honra).

De acordo com os resultados do estudo, foram encontrados os tipos de registro para a universidade e os níveis de renda da família De acordo com os resultados do estudo dos alunos para ser associado com o sucesso do aluno. Isto mostra que os modelos têm capacidade de previsão. Os dados faltantes em algumas colunas no conjunto de dados influenciaram o sucesso do sistema diretamente. Portanto, se a quantidade de dados e o número de variáveis aumenta, previsões mais precisas sobre o sucesso do aluno podem ser realizadas.

Tsai *et al.* (2010) investigaram uma nova abordagem baseada em agrupamento de dados chamado de PHD (é um algoritmo baseado em densidade híbrida), é uma versão melhorada do KIDBSCAN. Assim sendo, os pares mais próximos do *cluster* são combinados até que o agrupamento do conjunto de dados seja alcançado. A finalidade do PHD é diminuir o custo e o tempo de execução, mostrando que o PHD é mais eficiente que o DBSCAN (*Density-based spatial clustering of applications with noise* - é um algoritmo de agrupamento densidade que é baseada num número de grupos começando com uma estimativa da distribuição da densidade dos nós correspondentes), IDBSCAN e KIDBSCAN em termos de tempo de execução.

Desta forma, os autores concluíram que o algoritmo PHD proposto pode ser utilizado de forma eficiente para o agrupamento de grandes conjuntos de dados, que o PHD é mais rápido do que KIDBSCAN, e que o número de consultas do algoritmo PHD proposto é menor do que KIDBSCAN com aproximadamente 65% a 85% em todas as simulações.

Pavanelli *et al.* (2011) utilizaram a RNAs e RLM com dois objetivos: criar uma "agenda inteligente", através da previsão do tempo de audiências e, também, possibilitar a "negociação" entre as partes, através da previsão da duração do trâmite de processos trabalhistas. Para tanto, utilizou-se os dados de processos trabalhistas do Fórum Trabalhista de São José dos Pinhas, PR, para fazer o treinamento de diversas RNAs com várias topologias e, também, da RLM. Os autores utilizaram a ACP e/ou a codificação dos atributos preliminarmente a utilização das referidas técnicas, visando melhorar ainda mais os seus desempenhos.

Vagin e Fomina (2011) consideraram o problema da informação de generalização e examinaram as formas de solução na presença de ruídos nos dados originais. Os ruídos podem aparecer devido a seguintes causas: medição incorreta dos parâmetros de entrada, descrição errada dos valores dos parâmetros, utilização de dispositivos de medição danificados, e perda de dados em transmissão e armazenamento da informação. Diversos modelos de ruído de informações são apresentados, e os autores discutiram a influência do ruído para os algoritmos de generalização.

Desta forma, foram considerados os modelos de ruído em tabelas de banco de dados relacionados com a ausência de valores de atributos ou a distorção e embaralhamento de valores de atributos em uma amostra de aprendizagem.

Os autores propuseram e descreveram o algoritmo IDTUV2 (*Induction of Decision Tree with restoring Unknown Values*/Indução de Árvores de Decisão com a restauração de valores desconhecidos) que permite o processamento de amostras de aprendizagem que contêm exemplos com valores de ruidosos. Os métodos utilizados foram: Árvores de Decisão e Regras de Produção.

Os resultados obtidos da modelagem mostram que os algoritmos C4.5 e CN2 em combinação com o algoritmo IDTUV2 permite restaurar a manipulação de dados de forma eficiente na presença de ruído de diferentes tipos.

Ngai *et al.* (2011) apresentaram uma ampla revisão de artigos acadêmicos e forneceram uma bibliografia abrangente e uma estrutura de classificação para as aplicações de DM para a FFD (Detecção de Fraudes Financeiras). Embora a FFD seja um tema emergente de grande importância, uma ampla revisão da literatura sobre o assunto ainda não foi realizada.

Foram selecionados 49 artigos de revistas entre 1997 e 2008. Na sequência, os autores analisaram e classificaram os artigos em 4 categorias de fraude financeira (fraude bancária, fraude de seguros, valores mobiliários e fraude mercadorias, e outras fraudes financeiras relacionadas) e em seis técnicas de DM (classificação, regressão, *clustering*, previsão, detecção de *outliers*, e visualização).

As principais técnicas de DM utilizadas para FFD são modelos logísticos, Redes Neurais, Rede Bayesiana, e Árvores de Decisão, os quais fornecem soluções primárias para os problemas inerentes à detecção e classificação de dados fraudulentos.

Os resultados mostram claramente que as técnicas de DM foram aplicadas mais extensivamente para a detecção de fraudes de seguros, embora as fraudes corporativas e fraudes de cartão de crédito também tenham atraído muita atenção nos últimos anos. Por outro lado, os autores encontraram uma clara falta de investigação sobre fraudes de hipoteca, lavagem de dinheiro e valores mobiliários e fraude de mercadorias. Este artigo também aborda as lacunas entre FFD e as necessidades da indústria para incentivar a pesquisa adicional sobre temas negligenciados.

Gagliardi (2011) estudou a viabilidade e o desempenho de alguns sistemas classificadores pertencentes à família de aprendizagem IB (*Instance-Based*/Baseada em Instância) como ferramentas de diagnóstico. O estudo foi realizado através de 3 bases de dados médicos: a 1ª. referente ao diagnóstico diferencial das doenças dermatológicas (eritemato-escamosas/*erythemato-squamous*); a 2ª. para o diagnóstico do desenvolvimento de diabetes mellitus e a 3ª. em um problema de diagnóstico por imagem em cardiologia nuclear. Foram aplicados em 5 classificadores IB em cada base de dados, duas baseadas em exemplares, uma baseada em protótipos e duas híbridas. A análise dos resultados experimentais mostra que os classificadores com os melhores desempenhos na classificação foram: o *k*-NNC (*k-Nearest Neighbour Classifier*/Classificador *k*-Vizinhos Mais Próximos), o NPC (*Nearest Prototype Classifier*/Classificador Vizinho Mais Próximo)

e o PEL-C (*Prototype Exemplar Learning Classifier*/Classificador de Aprendizagem por Exemplares). Os autores comentam que a complexidade e responsabilidade das práticas de diagnóstico requer que esses resultados sejam melhor confirmados em outros domínios clínicos.

Padhy, Mishra e Panigrahi (2012) analisaram diversas aplicações de DM. Os diferentes métodos de DM foram utilizados para extrair os atributos. A seleção de dados e os métodos para a extração de dados é uma tarefa importante neste processo e necessita do domínio do conhecimento.

Várias tentativas foram feitas para projetar e desenvolver o sistema de DM genérico, mas nenhum sistema encontrado foi completamente genérico. As aplicações de DM genéricos tiveram limitações. A partir do estudo de várias aplicações de exploração de dados, os autores observaram que nenhum aplicativo chamado de “genérico” é 100% genérico. As interfaces inteligentes e agentes inteligentes até certo ponto, fazem aplicações genéricas, mas têm limitações.

Os resultados de rendimento a partir do domínio de aplicações específicas são mais precisos e úteis. Deste modo, parece muito difícil projetar e desenvolver um sistema de DM que possa funcionar de forma dinâmica para qualquer domínio.

Kuo *et al.* (2012) propuseram uma nova abordagem de agrupamento dinâmico baseada na otimização por enxame de partículas (PSO) e AG através do algoritmo DCPG (*Distributed Computation Precedence Graph*/Computação Distribuída de Precedência Gráfica) para resolver o problema da fixação do número de *clusters* com antecedência e descobrir o número adequado de *clusters* com base nas características dos dados.

Quatro conjuntos de dados de referência com números diferentes de aglomerados, dimensões e tipos foram utilizados para verificar que o algoritmo DCPG pode obter o número certo de *clusters* e alcançar melhores resultados de *clustering*. Além disso, o algoritmo DCPG foi comparado com DCPSO, ACMPSO e DCGA. O algoritmo DCPG é a mais estável.

O algoritmo DCPG é aplicado para agrupar as listas de materiais (BOM) para a Companhia *Advantech* em Taiwan. Os resultados de agrupamento podem ser usados para categorizar produtos que partilham os mesmos materiais em *clusters*.

Zaragozí *et al.* (2012) estudaram os fatores que determinam o processo de abandono das terras agrícolas, combinando SIG (*Geographical Information Systems*/Sistemas de Informação Geográfica) e técnicas de KDD para definir as

variáveis mais importantes para o estudo do processo de abandono. As variáveis consideradas neste estudo podem ser agrupadas em ambientais, socioeconômicas e aquelas relacionadas a práticas agrícolas. Testes demonstraram que realizando a mesma análise para centenas de variáveis diferentes, é possível explorar novas relações que possam ser úteis no estudo de um problema complexo. Os autores mostraram a capacidade da técnica de DM (correlação entre os atributos) para selecionar as características mais importantes para a criação de cenários úteis. Os autores concluíram que a aplicação do processo KDD pode ser útil para a seleção das melhores combinações de variáveis em estudo no abandono de terra.

Para Erohin *et al.* (2012), a indústria de manufatura vem utilizando ferramentas digitais para o desenvolvimento de produtos e controle de produção para controlar o produto e a complexidade do processo, bem como para reagir à crescente pressão de custo e tempo.

Desta forma, a finalidade deste estudo foi a aplicação do processo de KDD industriais para a identificação e extração de novos conhecimentos, a fim de apoiar os processos de planejamento e tomada de decisão no surgimento de produto.

Os autores descrevem as abordagens básicas para a utilização inteligente do conhecimento descoberto no exemplo de determinação prospectiva no tempo de montagem nas fases iniciais do surgimento de produto chamado de PEP (*Product Emergence Process*/Processo de Surgimento do Produto). Além disso, o conhecimento de planejamento que surge durante o PEP pode ser fornecido para todos os métodos aplicados e ferramentas (por exemplo, extensões de funcionalidades PLM (*Product Lifecycle Management*/Gerenciamento de Ciclo de Vida de Produto)). Ao mesmo tempo, os potenciais de KDD em manufatura digital são avaliados.

Yoo *et al.* (2012) discutiram a fundo a definição e o processo de DM, as principais diferenças entre a estatística e DM e a singularidade de DM em áreas de saúde e biomédicas. Os autores fizeram um breve resumo de vários algoritmos de DM utilizados para a classificação, *clustering* e associação, bem como de suas respectivas vantagens e desvantagens.

O DM tem sido amplamente utilizado nas áreas da saúde e biomédicas, por causa de seu poder descritivo e preditivo. Usando tecnologias de DM, os profissionais de saúde podem prever a fraude de seguros de saúde, pacientes subdiagnosticados, o custo da saúde, o prognóstico da doença, o diagnóstico da

doença, e o tempo de permanência em um hospital. Além disso, podem obter atributos frequentes de bases de dados biomédicos e de saúde, tais como as relações entre as condições de saúde e doença e as relações entre as drogas.

Desde modo, os autores concluíram que o DM tem vários problemas que impedem seu uso clínico por profissionais de saúde. Em primeiro lugar, os algoritmos de DM geralmente exigem parâmetro(s) de usuário(s), principalmente porque cada algoritmo de DM tem seus próprios pressupostos teóricos. Os usuários finais geralmente não têm informações suficientes sobre o(s) parâmetro(s) e sua seleção. Para piorar o problema, os resultados de DM são geralmente muito sensíveis ao(s) parâmetro(s).

Em segundo lugar, a precisão de DM não é, normalmente, suficientemente elevada para ser utilizado num ambiente clínico. Geralmente a baixa qualidade dos dados do paciente contribui para o problema, dado que os sistemas de informação hospitalar são normalmente concebidos para fins financeiros. Outra razão para a qualidade de dados ser baixa é que fatores biomédicos e de saúde que afetam as doenças não são totalmente conhecidos.

Por último, há uma completa falta de pacotes de DM completos para descoberta de conhecimento. Um pacote de DM ideal deve (1) apoiar o pré-processamento de dados inteligente que seleciona automaticamente e elimina os dados, e (2) automatizar completamente o processo de descoberta de conhecimento para que ele entenda e utilize conhecimento em processo de DM existente para melhor descoberta de conhecimento.

Os autores acreditam que se esses problemas forem devidamente resolvidos, o DM pode se tornar um núcleo de tecnologia necessário para a prática da medicina baseada em evidências.

Liao, Chu e Hsiao (2012) apresentaram uma revisão da literatura relacionada com as aplicações e técnicas de DM, no período de 2000 a 2011, a fim de determinar como essas técnicas e aplicações têm se desenvolvido.

Para isto, utilizaram os índices de palavras-chave e resumos de artigos para selecionar os mesmos através de cinco bases de dados *on-line* e 159 revistas acadêmicas, resultando em 216 artigos. Os autores analisaram e classificaram a técnica de DM em relação as três seguintes áreas: tipos de conhecimento, tipos de análise e tipos de arquitetura, juntamente com suas aplicações em diferentes domínios e práticas.

Os resultados mostraram que diferentes metodologias das ciências sociais, como a psicologia, a ciência cognitiva e o comportamento humano podem implementar a técnica de DM, como uma metodologia alternativa. A capacidade de mudar continuamente e proporcionar um novo entendimento são as principais vantagens de DM.

Carmona *et al.* (2012) apresentam um estudo experimental relativo ao efeito da utilização de tratamento de dados faltante em EFSS (*Evolutionary Fuzzy Systems*/Sistemas Nebulosos Evolutivos) para a SD (*Subgroup Discovery*/Descoberta de Subgrupos), onde as abordagens de imputação mais relevantes para o tratamento de MVs (*Missing Values*/Valores Faltantes) são utilizadas, a fim de pré-processar alguns conjuntos de dados padrão.

Foi realizado um estudo experimental com um número de conjuntos de dados que contêm ambos os MVs naturais e induzidos.

Os resultados mostraram que, entre os métodos estudados, a abordagem de pré-processamento KNNI (*Imputation with K-Nearest Neighbor*/Imputação com K Vizinhos Mais Próximos) para MVs obtém os melhores resultados em EFSS para a SD.

Bina *et al.* (2013) apresentam uma nova maneira de aprimoramento para a classificação de dados relacional (*Relational Data Classification* - fazer previsões não só das tabelas, mas também dos objetos relacionados). Desta forma, o intuito foi estudar as diferentes maneiras independentes de Árvores de Decisão através de tabelas de bancos de dados, e em seguida, criar um modelo Log-Linear para prever as probabilidades de classe. Neste estudo, foi utilizada a RL como o classificador probabilístico de base, ao invés de Árvores de Decisão.

Os resultados demonstraram um melhor desempenho preditivo nos tempos de execução através da avaliação empírica em três conjuntos de dados.

Assim sendo, este método teve por finalidade mostrar as características que difere dos outros métodos, que são as seguintes: (1) As funções de agregação não são utilizadas, o que evita a perda de algumas informações e permite uma aprendizagem eficiente; (2) As informações de todas as ligações são consideradas na classificação; (3) A RL é usada para ponderar informações de diferentes tabelas.

Tripathy *et al.* (2013) conduziram um experimento na região semi-árida da Índia para compreender as relações de integração doenças/*crop-weather-pest* (colheita-clima-praga) usando *wireless* sensorial e dados de vigilância em nível de

campo sobre as pragas estreitamente ligadas a pragas interdependentes (*Thrips*, que são insetos pequenos); doença (*Bud Necrosis*, que é uma das principais doenças do amendoim), ou seja, na dinâmica de colheita do amendoim (*dynamics of groundnut (peanut) crop*).

Várias técnicas de DM foram utilizadas para transformar os dados em informações úteis, conhecimento, relações, tendências e correlação de colheita-clima-praga e evolução das doenças. Essas dinâmicas obtidas através das técnicas de DM e treinadas por meio de modelos matemáticos foram validadas com os dados de vigilância do nível do solo correspondente.

Deste modo, os autores constataram que a infecção da doença viral *Bud Necrosis* é fortemente influenciada pela umidade, a temperatura máxima, a duração prolongada da umidade da folha (folha molhada), a idade da colheita e por uma praga transmissora a *Thrips*.

A abordagem estatística juntamente com a mineração ajudou no desenvolvimento do modelo de regressão multivariada, e tem sido utilizada para desenvolver um modelo de previsão empírica (não cumulativo) para emitir a previsão para o acúmulo de população, a iniciação e a gravidade da praga/doença.

No entanto, esta é uma investigação preliminar limitada a uma data de semeadura e isso tem de ser experimentado/validado continuamente com diferentes datas de semeadura, com experimentos de longa duração. Um modelo de previsão cumulativo foi desenvolvido para ajudar no desenvolvimento do Sistema de Apoio à Decisão para predição de praga/doença. Consequentemente ajudou a tomar decisões estratégicas, de modo a salvar a lavoura de pragas/doenças que afetam e melhorar o rendimento da colheita e das condições ambientais.

Ioannou *et al.* (2013) propuseram uma técnica de DM eficaz para a análise de dados biológicos e biomédicos. O processo de mineração proposto foi eficaz o suficiente para ser aplicado a vários tipos de dados biológicos e biomédicos. Para provar o conceito, foi aplicada a técnica de DM em duas áreas distintas, incluindo documentos de texto e dados biomédicos. Além disso, com base na abordagem proposta, os autores desenvolveram duas ferramentas de mineração, a *Bio Search Engine* e o *Genome-Based Population Clustering*.

No entanto, os autores pretendem expandir a ferramenta de mineração de texto biomédico vigente com capacidade avançada através da integração de técnicas de análise de usuário. Para a primeira versão da ferramenta *Bio Search*

Engine, pretendem criar um aplicativo com janelas independentes. Para o *Genome-Based Population Clustering*, pretendem criar técnicas de análise de DM mais alternativas. Além disso, pretendem combinar as duas representações diferentes de populações de uma forma mais sofisticada com o objetivo de melhorar a eficácia do método de agrupamento.

Sim *et al.* (2013) apresentaram os problemas de agregação, as definições de *cluster*, os algoritmos de agrupamento, o agrupamento subespaço de base e os trabalhos relacionados em *cluster* de alta-dimensionalidade.

O agrupamento subespaço inicial concentra-se em grupos de mineração através do conjunto de dados de alta dimensão, onde os objetos de um *cluster* estão fechados juntos em um subespaço do conjunto de dados.

Devido à proliferação de dados e do avanço da coleta de dados e a necessidade de resolver as tarefas mais complexas e exigentes, a pesquisa utilizou a aglomeração de subespaço cujo foco é: (1) a manipulação de dados complexos, como dados 3D, dados categóricos, fluxo de dados ou dados ruidosos; e (2) melhorar os resultados de *clustering*.

Assim sendo, já existe um amplo avanço sobre este tema, mais ainda há muitos problemas em aberto que devem ser discutidos.

Diamantini, Potena e Storti (2013) propuseram uma abordagem semântica orientada a serviços para o desenvolvimento de uma plataforma para a partilha e reutilização de recursos (processamento de dados e técnicas de mineração), permitindo a gestão de diferentes implementações da mesma técnica, com todas as funcionalidades para a produção e consumo de recursos, bem como provedores de recursos com diferentes capacidades técnicas e de domínio específicos.

A estrutura e a plataforma KDDVM (*Knowledge Discovery in Databases Virtual Mart*) foram concebidas para satisfazer a flexibilidade, a transparência, a confiabilidade e os requisitos de facilidade de uso. Além disso, a representação semântica de dados e serviços permitiu aos autores apresentar em KDDVM com um conjunto de serviços de apoio que são concebidos para dar suporte avançado ao esconder detalhes técnicos, assim satisfazendo a facilidade de uso e requisitos de transparência.

A utilização da plataforma foi proposta a um grupo de estudantes de PG em ciência da computação, durante o seu primeiro curso de KDD. Os *feedbacks* obtidos foram positivos, uma vez que melhora a precisão e o recall de resultados da

pesquisa. Ainda, os usuários têm apreciado muito a aceleração dada pelo *ClientFactory* e *KDDDesigner* para a definição de um experimento, já que evitam erros e fazem sugestões no ajuste de parâmetros e composição do processo, enquanto fornece uma interface única para qualquer serviço.

Kamsu-Foguem, Rigal e Mauget (2013) observaram que profissionais e acadêmicos têm um interesse comum no desenvolvimento contínuo de métodos de programas computacionais que suportam ou executam tarefas de engenharia de conhecimento intensivo para diminuir os tempos e custos de produção, pois são problemas que afetam diretamente o desempenho e a qualidade dos sistemas industriais. Desta forma, a Associação de Regras se apresenta como uma técnica de DM usada para descobrir informações úteis e valiosas em grandes bases de dados.

Os autores desenvolveram uma base conceitual para melhorar a aplicação do método de mineração através de Associação de Regras, para extrair conhecimento em operações e gestão da informação. A ênfase do estudo é a melhoria dos processos de operações. Assim sendo, analisaram o processo de manufatura de um provedor totalmente integrado de produtos de perfuração, a fim de realmente compreender as possibilidades e limitações da abordagem. Os resultados do experimento sobre conjuntos de dados reais mostram que a abordagem proposta é útil na busca de conhecimento efetivo associado a disfunções causais (*Dysfunctions Causes*). Ao mesmo tempo, o sistema ainda precisa de outras ferramentas de processamento.

Orriols-Puig *et al*. (2013) apresentam um estudo de caso utilizando um sistema inteligente novo que incorpora Lógica *Fuzzy* e AGs para operar de forma não supervisionada para extrair Regras de Associação *Fuzzy* de um conjunto de exemplos. Esta abordagem permite a Descoberta de Regras de Associação interessantes, que podem ser linguisticamente interpretadas, em bases de dados de grande escala (KDD). O objetivo dos experimentos foi a aplicação em um problema de canal de distribuição, com os seguintes objetivos: (1) estudar a robustez do sistema proposto para lidar com dados de *marketing*, (2) examinar a capacidade do sistema para suportar as conclusões obtidas pela aplicação da teoria direcionada à abordagem, (3) mostrar a capacidade do sistema para descobrir novas relações entre as variáveis e especificar intervalos das variáveis, que possam ajudar a entender melhor as verdadeiras associações entre as construções do problema, e (4) demonstrar a facilidade de utilização do sistema.

Com base nos resultados, pode-se dizer que o sistema produziu conclusões semelhantes quanto à abordagem teórica em primeira instância. Além de confirmar a hipótese inicial, o sistema proposto forneceu uma explicação de que cada hipótese foi apoiada ou rejeitada. Como o sistema não depende de qualquer modelo a priori pode-se descobrir novas associações interessantes entre duas ou mais variáveis que integram determinado cenário. O uso da Lógica *Fuzzy* resultou em um tipo de regras de associação que pode ser facilmente lido por especialistas do assunto, e também por profissionais de gestão, especialmente quando comparado com os valores probabilísticos obtidos pelos testes estatísticos usualmente aplicados pela academia tradicional.

Desta forma, a principal característica do sistema é que ele pode descobrir automaticamente novas associações interessantes diretamente dos dados. O sistema pode ser especialmente útil para descobrir as relações negligenciadas por especialistas em *marketing* em situações como uma estrutura inadequada ou delimitação do problema ou a falta de informação a priori para orientar a análise original.

Soto *et al*. (2013) utilizaram um processo iterativo baseado em KDD para obter uma melhor descrição da atividade microbiana de ferro e a taxa de dissolução de minérios de sulfeto que ocorrem no ciclo de lixiviação. Foram utilizadas duas técnicas de DM: Agrupamento Hierárquico e Árvore de Decisão, através do *PASW Statistics* 18 *Software* (SPSS) e o Algoritmo de CRT ou CART (*Classification and Regression Trees*). O Agrupamento Hierárquico foi realizado para descobrir um conjunto de grupos, utilizando dados mineralógicos (normalizados entre 0 e 1), distância euclidiana e o método de *Ward*. Na sequência gerou as Árvores de Decisão formando três grupos de diferentes faixas. Deste modo, usando os dados provenientes da análise dos minerais despejados nas diferentes faixas de acúmulo industrial, foi possível concluir a relação entre a disponibilidade do substrato mineral e o crescimento microbiano.

Uma proposta de representação gráfica e visualização exploratória para Árvores de Decisão no processo KDD, especificamente na etapa de DM, através de uma técnica simples de tabelamento de partição, baseada na técnica *Treemap* (*maps of trees*/mapas de árvores), que permite representar estruturas hierárquicas foi apresentada por Rojas e Villegas (2013). O estudo avalia e compara a técnica por meio de quatro critérios: eficiência visual, informação de alto/baixo nível, tipo de

dado e a expressividade. Os resultados mostraram que a técnica é capaz de descrever as características da classe, para facilitar a comparação entre diferentes objetos, permitindo implementar facilmente itens de dados, onde os dados são os originais e não os padronizados, podendo o usuário alterar as variáveis usadas para construir o gráfico. Com isto pode-se visualizar claramente as pequenas árvores em duas e três dimensões e entender com facilidade a distribuição de nós em cada nível.

Nieminen, Pölönen e Sipola (2013) aplicaram o processo de descoberta do conhecimento para o mapeamento dos temas atuais em um determinado campo da ciência. O interesse é em saber como os artigos formam aglomerados e quais são os conteúdos dos grupos encontrados.

Os autores criaram um *framework* de mapeamento da literatura com base no aglomerado de artigos publicados em revistas de alto impacto. Deste modo, propuseram a análise de um estudo de caso ocorrido em 2011.

A seleção e interpretação dos dados foram realizadas de maneira manual. A metodologia foi totalmente automatizada e os passos individuais poderiam ser alterados, se um método mais adequado fosse descoberto. Por causa da automação do processo, o estudo é menos tendencioso do que as pesquisas que utilizam a abordagem baseada na opinião.

A metodologia deve ser útil para os indivíduos e as empresas que tentam ganhar uma compreensão de grandes conjuntos de dados textuais, por exemplo, documentação interna pessoal ou da empresa. Deve ser útil também para os cientistas do campo da aplicação e as empresas que querem encontrar métodos que são usados atualmente.

Atualmente, a “saída” (produção) do método é um instantâneo dos atuais artigos publicados. Combinando com um ponto de vista longitudinal pode-se revelar tendências de longo prazo na literatura de pesquisa. A abordagem poderia se beneficiar de informação adicional adquirida com recursos extraídos de resumos. Os resumos estão geralmente disponíveis, além de palavras-chave e títulos, enquanto que outras partes dos artigos podem não estar disponíveis.

Coussement, Van Den Bossche e De Bock (2014) investigaram o impacto do nível de precisão dos dados sobre o desempenho dos três algoritmos de segmentação: Análise de RFM (Recente, Frequência e Valor Monetário - é um método utilizado para a análise de valor do cliente), RLB e Árvores de Decisão. Os

autores utilizaram dois conjuntos de dados de *marketing* direto. O estudo recomenda a utilização de Árvores de Decisão no contexto de segmentação de clientes para *marketing* direto, mesmo sob a suspeita de problemas de precisão dos dados.

Pomponio e Le Goc (2014) apresentam uma metodologia de Engenharia do Conhecimento chamada de TOM4D (*Timed Observation Modelling For Diagnosis* /Modelagem da Observação Cronometrada para Diagnóstico), onde propõem uma modelagem essencialmente orientada para a sintaxe em que o conteúdo semântico é introduzido de forma gradual e controlada através da abordagem conceitual CommonKADS, Padrão de trabalho da Lógica Formal e Padrão de trabalho do Tetraedro de Estados.

Os princípios e fundamentos do TOM4D foram apresentados juntamente com um exemplo didático que ilustra a abordagem de modelagem que, os autores acreditam, torna o conhecimento explícito dos especialistas de uma forma que este pode ser comparado com o processo real. Assim, esta metodologia permitiu construir um modelo de processo a partir do conhecimento dos especialistas e dados onde este modelo, por construção, pode ser diretamente associado ao conhecimento do modelo dos especialistas e, ao mesmo tempo, podem ser comparados com modelos de processos reais obtidos com base nos dados. Neste sentido, esta abordagem permitiu reduzir a distância entre o conhecimento e os dados dos especialistas, ligando os dois universos, propondo levar os dois campos de estudo no sentido de uma visão holística.

Kusakabe e Asakura (2014) desenvolveram uma metodologia de fusão de dados para estimar atributos comportamentais de viagens que utilizam dados de cartão inteligente para observar as mudanças contínuas de longo prazo nos atributos de viagens. O método pode ajudar os operadores de transportes a monitorar e obter características comportamentais dos viajantes observadas nos dados dos cartões inteligentes.

O método destina-se a melhorar a compreensão do comportamento dos viajantes durante o monitoramento dos dados de cartões inteligentes. A fim de completar atributos comportamentais ausentes nos dados de cartão inteligente, este estudo desenvolveu uma metodologia de fusão de dados de cartões inteligentes com os dados do levantamento sobre viagem da pessoa com o modelo probabilístico *Naive Bayes*. O resultado da validação mostrou que o método proposto estima com sucesso os fins de viagem em 86,2% dos dados de validação.

A análise de DM empíricos mostrou que a metodologia proposta pode ser aplicada para encontrar e interpretar as características comportamentais observadas nos dados de cartão inteligente.

Com o objetivo de melhorar o desempenho operacional de um edifício mais alto de Hong Kong, Xiao e Fan (2014) investigaram o uso de DM para a análise de grandes conjuntos de dados em um BAS (*Building Automation System*/Sistema de Automação Predial). A preparação de dados foi realizada para melhorar a qualidade dos dados e transformá-los em formato adequado para a DM. A análise de agrupamento foi realizada para identificar os atributos de operações para controle de temperatura no edifício. A Associação de Regras foi realizada para desvendar as associações entre os consumos de temperatura dos principais componentes em cada *cluster* (grupo). A pós-mineração foi realizada para selecionar e interpretar as possíveis regras úteis. Dois casos foram apresentados para demostrar a aplicabilidade em melhorar o desempenho operacional do edifício.

Spruit, Vroon e Batenburg (2014) basearam sua pesquisa no modelo CRISP-DM (*CRoss Industry Standard Process for Data Mining*/Processo Padrão Inter Indústrias para DM) para estruturar o processo de descoberta de conhecimento aplicado em um estudo de caso exploratório em uma instituição Holandesa de cuidados holandês de longo prazo (*Dutch long-term care institution*). A coleta de dados foi feita por meio de entrevista com 22 especialistas de Casa de Repouso, Lares e Casa de Cuidados (*Care homes and Home Cares*), Cuidado em Saúde Mental (*Mental care*), e Cuidado Deficiência (*Disability Care*) para determinar as necessidades das informações.

Estas informações foram traduzidas em 25 metas de DM quantificáveis e selecionada a instituição com cerca de 850 clientes em 5 locais. Na sequência, analisaram o banco de dados da instituição Holandesa de cuidados holandês de longo prazo que continham informações de 2008 a 2012 para identificar atributos em informações de incidentes, atributos de informação de avaliação de risco, a relação entre as avaliações de risco e informações de incidentes, atributos na duração média da estadia, e identificar e prevenir a *Care Intensity Package* (ZZP).

Consequentemente posicionaram todas as metas de DM em uma matriz da ordem de 2x2 para visualizar a importância relativa de cada objetivo em relação a qualidade do atendimento e a situação financeira das instituições de cuidados. Os autores concluem que explorando técnicas de descoberta de conhecimento com

base em um conjunto de dados e guiados por informações representativas precisa, pode-se contribuir para uma melhora na qualidade dos cuidados e nos gastos financeiros.

Karray, Chebel-Morello e Zerhouni (2014) estudaram o aspecto dinâmico do processo e serviço em uma plataforma de manutenção, para atender as necessidades crescentes no campo da manutenção. Os autores propuseram uma experiência de abordagem Feedback dinâmica, para explorar comportamentos do processo de manutenção na execução real da plataforma de manutenção.

Os autores utilizaram um sistema baseado em rastreamento chamado de "PETRA" (*trace-based system*). O sistema é composto pelo monitoramento, aprendizagem e capitalização do conhecimento. Para simular este sistema foi utilizado o jBPM6 e a plataforma *Weka*. Os resultados de aprendizagem são explorados pelo regime de capitalização para validar as regras de confiança e para alimentar a base de conhecimento. Os resultados obtidos do processo de manutenção mostraram que os ativos (habilitados) podem ser atualizados e também garantir que o retorno de experiência seja adaptado às necessidades do usuário.

Deste modo, a finalidade do estudo foi extrair regras de conhecimento das atividades da plataforma e as atuações do usuário (ou seja, a experiência de operadores de manutenção). A experiência coletiva foi explicitada através das regras de descobertas a partir da resolução de problemas repetitivos decorrentes de atividades de manutenção. Este conhecimento é específico para o comportamento de cada sistema mantido, e permite que os utilizadores resolvam os problemas, sem olhar para as soluções. De fato, o conhecimento permite a evolução dinâmica dos processos de manutenção e, assim, permite a atualização geral.

Qi *et al.* (2014) propuseram um novo algoritmo paralelo, o PRMCLP (*Parallel Regularized Multiple-Criteria Linear Programming*) para superar as exigências da computação e do armazenamento que aumentou rapidamente com o número de amostras de treinamento. Primeiramente, o modelo RMCLP (*Regularized Multiple-Criteria Linear Programming*) foi convertido em um problema de otimização sem restrições, e depois dividido em várias partes, sendo que cada parte foi calculada por um único processador. Os autores analisaram o resultado de cada uma das partes e ao fazer isso, obtiveram uma solução final de otimização de todo o problema de classificação. Com a ajuda de vários processadores, o desempenho do PRMCLP através dos conjuntos de dados melhorou consideravelmente.

Zhuk, Ignatov e Konstantinova (2014) propuseram extensões do clássico método JSM (*John Stuart Mill*) e do classificador *Naïve Bayesian* para o caso de dados relacionais triádicos. Foi realizada uma série de experiências em vários tipos de dados (reais ou sintéticos) para estimar a qualidade da técnica de classificação e compará-la com outros algoritmos de classificação que geram hipóteses como, por exemplo, ID3 e *Random Forest* (Florestas aleatórias). Além da precisão da classificação, também avaliaram o desempenho do tempo dos métodos propostos.

O método JSM mostrou resultados relativamente bons, enquanto que o método original JSM, mesmo tendo altos valores de medida F, deixou uma grande fração de exemplos não classificados. Deste modo, devido às peculiaridades dos dados reais *Bibsonomy*, todos os métodos de classificação não foram satisfatórios. Foram parcialmente superados usando meta-informações como condições formais adicionais. Para estudos futuros, os autores pretendem considerar técnicas de classificação mais flexíveis baseadas em ACO-*triclusters*.

Engel *et al.* (2014) consideram que as ferramentas de DM podem ser computacionalmente exigentes, por isso há um interesse crescente sobre as estratégias de computação paralela para melhorar seu desempenho. A população de GPUs (*Graphics Processing Units*/Unidades de Processamento Gráfico) aumentou o poder de computação de desktops, mas as ferramentas de DM baseadas em desktops não costumam aproveitar o máximo dessas arquiteturas. A popularização de GPUs representa novas oportunidades para a paralelização de *software* e melhoria de desempenho acessível para o usuário final sobre o seu próprio *desktop*. DM é um dos campos de aplicação que podem se beneficiar desses avanços.

Deste modo, foi proposta uma abordagem para melhorar o desempenho do *Weka*, por meio da paralelização em máquinas aceleradas por GPU. Os autores optaram por paralelizar um método de multiplicação de matrizes usando a ferramenta *state-of-the-art* (estado da arte). Os autores identificaram um conjunto de operações que podem ser facilmente adaptadas para GPUs através do uso do *Java profilers*. Como resultado, observaram-se níveis de velocidade de pelo menos 49% em relação aos resultados anteriores, diminuindo drasticamente o tempo que o algoritmo consome para lidar com um conjunto de dados. Tem-se, assim, um aumento significativo de velocidade com as arquiteturas paralelas, em comparação ao original, com código *Weka* sequencial.

Relich e Muszynski (2014) investigaram o uso de sistemas inteligentes para identificar os fatores que influenciam significativamente o tempo de desenvolvimento de novos produtos. Neste estudo vários modelos e metodologias do processo de descoberta de conhecimento foram comparados e definidos por 4 etapas: seleção de dados, transformação de dados, DM e interpretação dos resultados. Entre as técnicas de DM, duas foram escolhidas: as RNA e o Sistema *Neuro-Fuzzy*. Estas técnicas foram escolhidas para buscar relações entre a duração da fase do projeto e outros dados armazenados no sistema de informação da empresa. A análise dos resultados mostra que os sistemas inteligentes têm uma melhor qualidade de estimativa em relação os modelos estatísticos e que a seleção das variáveis no pré-processamento influencia significativamente os resultados obtidos.

Luque-Baena *et al.* (2014) analisaram três conjuntos de dados de câncer (leucemia, câncer de pulmão e câncer de próstata) utilizando uma abordagem combinada de AGs e informações biológicas extraída do banco de dados KEGG (*Kyoto Encyclopedia of Genes and Genomes*), a fim de obter uma seleção robusta de recurso de subconjunto com bons índices de desempenho. A abordagem incorpora um novo método de recurso de pontuação dentro do AG, levando em conta as informações biológicas sobre proteínas (principalmente enzimas) envolvidas nas vias dos transtornos estudados.

A descoberta mais notável é que a proposta melhora a estratégia padrão do AG independentemente do modelo de classificação utilizado (LDA ou SVM) nos três conjuntos de dados analisados, obtendo resultados estatisticamente significativos em dois deles (Leucemia e pulmonar). Os resultados mostraram a importância da informação biológica sobre o diagnóstico de câncer (ou para qualquer outra doença). Portanto, esta abordagem poderá facilitar a definição de perfil gênico para o prognóstico clínico e diagnóstico de doenças cancerosas. Além disso, poderá também ser utilizada para a descoberta de conhecimento biológico sobre a doença em estudo.

Hasumi e Kamioka (2014) propuseram um sistema de previsão de aplicação que recomenda um aplicativo para o usuário quando está usando o computador para algum trabalho, através da RNAs. Além disso, a eficácia do sistema proposto foi discutida mostrando a precisão da previsão de cerca de 90% em recomendar aplicações úteis, quando o usuário utilizar o computador no dia a dia.

Holm, Korman e Ekstedt (2015) apresentaram um modelo baseado em rede Bayesiana que pode ser utilizado pelos tomadores de decisão de empresa para estimar a probabilidade de que um testador de penetração profissional (*professional penetration tester*) que é capaz de obter informação sobre as vulnerabilidades críticas e *exploits* (segurança de computadores) para *software* em diferentes circunstâncias.

Os dados foram obtidos a partir de estudos empíricos anteriores por meio dos bancos de dados *on-line* e de uma pesquisa com 58 indivíduos onde todos foram creditados para a descoberta de vulnerabilidades de *softwares* críticos. O modelo proposto descreve 13 estados relacionados por 17 atividades, e um total de 33 conjuntos de dados diferentes. Os autores concluem que as estimativas do modelo podem ser usadas para apoiar as decisões sobre qual *software* deve ser adquirido, ou quais são as medidas para investir em projetos de desenvolvimento de *softwares*.

Os trabalhos correlatos acima descritos podem ser resumidos de acordo com o Quadro 3.2, a seguir, onde são apresentados os autores do trabalho, o ano em que o mesmo foi desenvolvido, o local, país e estado, onde foram aplicados, quais as técnicas utilizadas (na análise exploratória dos dados e em DM) e quais foram os resultados obtidos.

Quadro 3.2 - Resumo dos trabalhos correlatos

Autores/Ano	**Local da Aplicação**	**Técnicas preliminares**	**Técnicas de DM aplicadas**	**Resultados (promissores ou não) e *softwares* (prontos ou desenvolvidos)**
Steiner; Carnieri (1999)	Agência bancária de Curitiba, PR, Brasil Registros históricos de clientes (pessoa física)	Correlação entre os dados e T^2 de Hotelling	RL	Promissor - *software* estatístico GLIM
Lemos; Steiner; Nievola (2005)	Agência bancária de Guarapuava, PR, Brasil Análise de registros históricos de clientes (pessoas jurídicas) de uma agência bancária	Binarização dos Dados	Redes Neurais, Árvore de Decisão	Promissor - *software WEKA e o software MATLAB*
Steiner; Soma; Shimizu; Nievola; Steiner Neto (2006)	Hospital das Clínicas de Curitiba, PR, Brasil Problema médico	T^2 de Hotelling, transformação dos atributos e descarte de pontos atípicos	GSME-PL, PL, FDLF, RL, Redes Neurais e Árvores de Decisão.	Promissor – programa computacional Visual Basic, *software* Lingo, *software* estatístico GLIM e *software WEKA*

(“continua”)

(“continuação”)

Steiner; Nievola; Soma; Shimizu; Steiner Neto (2007)	Agência bancária de Guarapuava, PR, Brasil Crédito bancário	Codificação dos atributos (Termômetro e *Dummy*)	Algoritmo *NeuroRule*	Promissor – *software WEKA*
Baptistella; Cunico; Steiner (2007)	Prefeitura Municipal de Guarapuava, PR Análise imóveis residenciais (casas e apartamentos)	ACP	RNAs	Promissor - *software Statística* e o *software MATLAB*
Steiner; Chaves Neto; Braulio; Alves (2008)	Conjunto de imóveis, da cidade de Campo Mourão, PR. Imóveis urbanos nas classes de apartamentos, residências e terrenos	Escore de discriminação quadrático	Análise de agrupamentos e RLM	Promissor
Steiner; Carnieri; Stange (2009)	Indústria de papel paranaense	T^2 de Hotelling e Distância de *Mahalanobis*	Dois modelos de PL; Fisher; k-vizinhos mais próximos e RLB	Promissor - linguagem Pascal utilizando o pacote computacional GAMS e GLIM
Bettiollo Junior; Steiner (2009)	Dados médicos – mulheres grávidas	Codificação dos atributos	RNAs e RL	Promissor - *software Statgraphics*
Martínez-López; Casillas (2009)	Comportamento dos consumidores em ambientes *on-line*	Transformar a escala de intervalo em semântica *Fuzzy*	Sistemas *Fuzzy* Genéticos	Promissor - *software* em desenvolvimento
Fang; Rachamadugu (2009)	Empresa de *Marketing* na Internet - *ComScore* Media *Matrix* Registros de consumidores *on-line*		Programação Dinâmica	Promissor
Mendes; Fiuza; Steiner (2010)	Clínica neurológica	Binarização dos Dados	RNAs	Promissor - *software MATLAB* e o componente *Neural Network Toolbox*
Tsai; yeh; chang; liu (2010)	Banco de dados da Iris e de reconhecimento de vinho	Análise de Cluster	Agrupamento de Dados K-*means* e PHD	Promissor
Guruler; istanbullu; karahasan (2010)	Faculdade de Economia e Ciências Sociais na Universidade de Mugla a partir de 1995 Dados demográficos dos estudantes universitários	Codificação dos atributos	Árvore de Decisão	Promissor - *software* de descoberta de conhecimento, MUSKUP
Pavanelli; Pavanelli; Steiner; Costa; Gusmão (2011)	Processos trabalhistas (Fórum de São José dos Pinhais, PR)	Codificação dos atributos e ACP	RNAs e RLM	Promissor

(“continua”)

(“continuação”)

Vagin; Fomina (2011)	Conjuntos de dados de teste da Universidade de Informática e Engenharia de Computação da Califórnia por meio da UCI Informação de generalização	Codificação dos atributos	Árvores de Decisão e Regras de Produção	Promissor - algoritmo IDTUV2
Gagliardi (2011)	Dados médicos	IB	K-NNC, NPC e PEL-C	Promissor
Ngai; Hu; Wong; Chen; Sun (2011)	Banco de dados de revistas acadêmicas desde 1997 a 2008	Estrutura gráfica de classificação	Modelos Logísticos, Redes Neurais, Rede Bayesiana e Árvores de Decisão	Promissor
Zaragozí; Rabasa; Rodríguez-Sala; Navarro; Belda; Ramóna (2012)	Terras agrícolas Dados climáticos (temperatura e precipitação de dados entre 1950 e 2007) – nível de estação meteorológica do Instituto Nacional de Meteorologia (INM)	As parcelas abandonadas	SIG e KDD	Promissor - algoritmo RBS
Erohin; Kuhlang; Schallow; Deuse (2012)	Linha de montagem		PLM e KDD	Promissor
Yoo; Alafaireet; Marinov; Pena-Hernandez; Gopidi; Chang; Hua (2012)	Áreas da saúde e biomédicas		Classificação, *Clustering* e Associação de Regra	Não promissor
Padhy; Mishra; Panigrahi (2012)	Análise das diversas aplicações de DM	-	-	Não promissor
Kuo; Syu; Chen; Tien (2012)	Departamento de Informação e Ciência da Computação através UCI Base de dados da Íris, Vinho, Vidro e Conjunto de dados de vogais	Normalização dos dados	Agrupamento dinâmico e AG.	Promissor algoritmo DCPG
Liao; Chu; Hsiao (2012)	Bancos de dados on-line desde 2000 a 2011 Revisão da literatura	-	-	Promissor
Carmona; Luengo; González; Jesus (2012)	Base de dados KEEL, Iris, Pima, novo-Tiróide, Ecoli, alemão, mágica e transporte	MVs	EFSS e SD	Promissor
Bina; Schulte; Crawford; Qian; Xiong (2013)	Banco de dados financeiros, hepatite, mondial, filmes e JMDB		RL e Árvore de Decisão	Promissor
Orriols-Puig; Martínez-López; Casillas; Lee (2013)	Canal de distribuição		Lógica Fuzzy, AGs e Regras de Associação Fuzzy.	Promissor

(“continua”)

(“continuação”)

Soto; Galleguillos; Serón; Zepeda; Demergasso; Pinilla (2013)	Industria de Minério localizada a 170 Km ao Sudeste de Antofagasta Ciclo de lixiviação	Agrupamento hierárquico (normalizados entre 0 e 1), distância Euclidiana e o Método Ward	Classificação e Regressão por árvores CART ou CRT	Não Promissor - PASW *Statistics* 18 *Software* (SPSS) e Algoritmo de CRT ou CART
Rojas; Villegas (2013)	Diversos estudos na área de computação em relação a comparação das técnicas	Visualização exploratória	Árvores de Decisão no processo KDD e *Treemap*	Promissor
Kamsu-Foguem; Rigal; Mauget (2013)	O exemplo de fabricação ocorre na Vam Drilling (uma parte da divisão de petróleo e gás da Vallourec & Mannesmann Tubes, que é uma filial do Grupo Vallourec) em Tarbes (Sudoeste da França) Produtos de perfuração	Codificação dos atributos	Associação de Regra	Promissor - algoritmo *RuleGrowth*
Nieminen; Pölönen; Sipola (2013)	Banco de dados das revistas Aglomerado de artigos - clustering	Classificação dos artigos usando fatores de impacto	Análise de Agrupamento	Promissor
Tripathy; Adinarayana; Sudharsan; Vijayalakshmi; Merchant; Desai (2013)	Instituto de Pesquisa Agrícola na Universidade Agrícola, Hyderabad na região Semi-árida da Índia Área agrícola (pragas/doenças)		Técnicas de DM e Regressão Multivariada	Promissor
Ioannou; Makris; Patrinos; Tzimas (2013)	Base de dados da Pubmed e da National Center for Biotechnology Information/Centro Nacional de Informação em Biotecnologia (NCBI) Dados biológicos e biomédicos	Análise de Cluster	*Bio Search Engine e Genome-Based Population Clustering.*	Promissor
Diamantini; Potena; Storti (2013)	Estudantes de PG em Ciência da Computação			Promissor - plataforma KDDVM
Sim; Gopalkrishnan; Zimek; Cong (2013)	Teórico/conceitual - Problema de agregação		Agrupamento Subespaço	Promissor
Pomponio; Goc (2014)	Linha de produção		TOM4D	Promissor
Kusakabe; Asakura (2014)	Comportamento dos viajantes		Naive Bayes	Promissor

(“continua”)

(“continuação”)

Xiao; Fan (2014)	Edifício mais alto de Hong Kong	Análise de Cluster	Análise de Agrupamento e Associação de Regra	Promissor
Spruit; Vroon; Batenburg (2014)	Instituição de cuidados holandês de longo prazo	Classificação dos dados através do software ECR	CRISP-DM	Promissor
Karray; Chebel-Morello; Zerhouni (2014)	Plataforma de manutenção		*Trace-Based System*	Promissor - *software Weka* e jBPM6
Qi; Alexandrov; Shi; Tian (2014)	Base de dados do MNIST (*Mixed National Institute of Standards and Technology*/Misto Instituto Nacional de Padrões e Tecnologia) e do UCI - repositório de aprendizado de máquinas Computação		RMCLP	Promissor - algoritmo PRMCLP
Roman; Ignatov; Konstantinova (2014)	Dados reais da bibsonomy Dados relacionais triádicos	Codificação dos atributos	JSM, *Naïve Bayesian*, ID3 e *Random Forest*	Não promissor
Engel; Charão; Kirsch-Pinheiro; Steffenel (2014)	Base de dados do UCI Máquinas de computação - GPUs		DM	Promissor - *software Weka*
Relich; Muszynski (2014)	Empresa de produção	ACP	RNAs e Sistema Neuro-Fuzzy	Promissor
Coussement; Van den Bossche; De Bock (2014)	Área empresarial - Dois conjuntos de dados de *marketing* direto	Estrutura de correlação	Análise de RFM, RL e Árvores de Decisão	Promissor
Luque-Baena; Urda; Claros; Franco; Jerez (2014)	Base de dados KEGG de leucemia, de próstata e de câncer de pulmão Dados médicos	Limpeza nos dados	AGs	Promissor
Hasumi; Kamioka (2014)	Registos de Log de Aplicação e Ontologia de Aplicação		RNAs,	Promissor
Holm; Korman; Ekstedt (2015)	Bancos de dados on-line e de uma pesquisa com 58 indivíduos	Combinação dos dados	Rede *Bayesiana*	Promissor

4 METODOLOGIA

Este capítulo tem por objetivo descrever como foi realizada a pesquisa e quais os instrumentos de apoio que foram utilizados para aplicar o processo KDD. Para poder comparar o desempenho dos métodos pesquisados quanto a classificação, três exemplos reais, de áreas distintas, descritos no capítulo 2, foram considerados: o problema nos cursos de PG Lato Sensu de uma IES privada, o problema do diagnóstico médico e o problema de calibração de balanças.

Para o problema nos cursos de PG Lato Sensu de uma IES privada os conjuntos de instâncias A e B representam as características resultantes das avaliações de alunos perante as disciplinas ministradas e qualidade dos serviços prestados, resultando em satisfação ou insatisfação com o curso.

Para o problema médico, os conjuntos de instâncias A e B representam as características de exames clínicos de pacientes, resultando em pacientes com câncer ou cálculo no duto biliar.

Para o problema de calibração de balança, os conjuntos de instâncias A e B representam as características envolvidas para a referida calibração, resultando em balanças calibradas ou não calibradas.

A metodologia proposta, aqui apresentada, foi dividida em duas fases, enquadradas no processo KDD (Figura 3.1 já apresentada). A 1ª. fase, que envolve a análise exploratória de dados, ficou composta pelas três técnicas: Detecção dos *Outliers* (Escore *Z* ou Distância de *Mahalanobis*), Teste T^2 de Hotelling, ACP apresentadas na seção 3.1.3 e a 2ª. fase, que envolve DM, ficou composta pelas cinco técnicas, RLB, GSME-PL, FDLF, RNAs e SVM apresentadas na seção 3.1.4, com o intuito de se obter a técnica com a máxima acurácia para o problema apresentado.

Resumidamente, a metodologia descrita acima pode ser visualizada na Figura 4.1, a seguir. O detalhamento será apresentado na sequência.

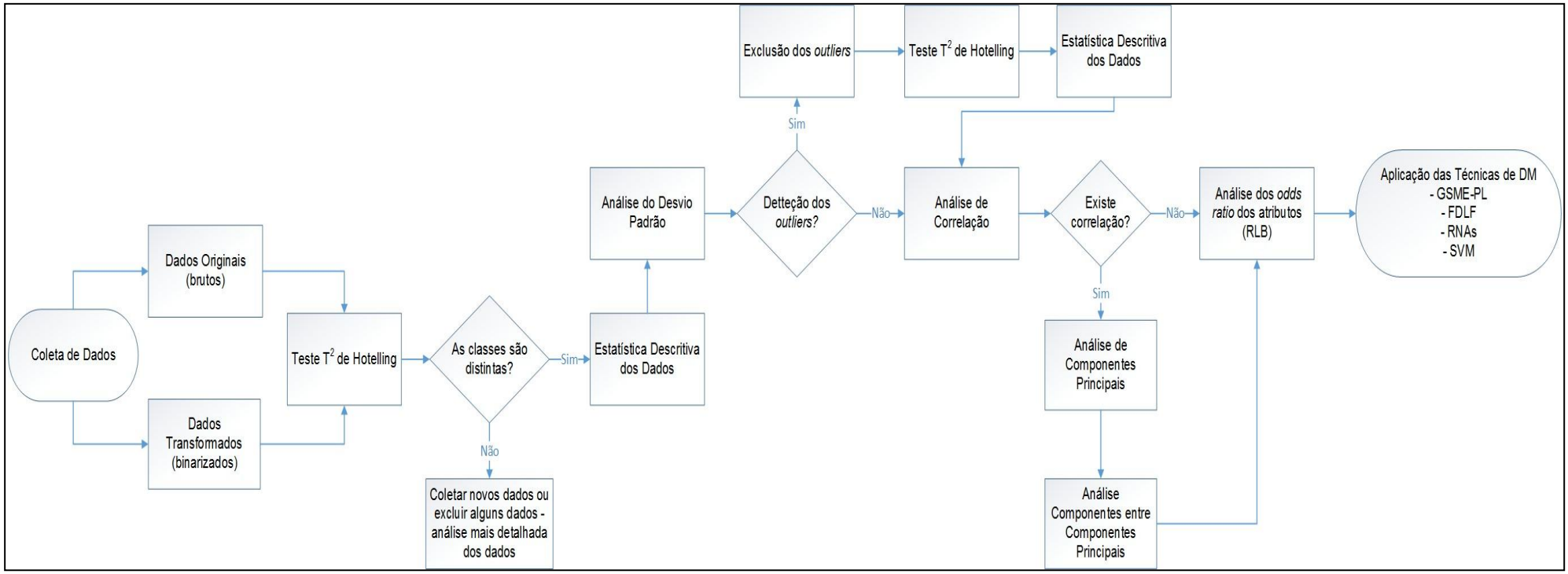

Figura 4.1 - Etapas da metodologia proposta
Fonte: Autoria própria

Para o problema dos cursos de PG Latu Sensu, o levantamento da base de dados foi realizado com o auxílio do software Excel, onde foram registradas 885 avaliações dos cursos de PG de uma IES privada. A instituição estudada disponibilizou os documentos, os quais foram digitalizados e transcritos para a planilha com o auxílio do software Excel (ANEXO A).

Já para o problema médico, foram registrados 118 pacientes do HC. A empresa estudada disponibilizou os documentos, os quais foram digitalizados e transcritos para a planilha com o auxílio de um médico especialista da área.

Para o problema de calibração de balança, o levantamento da base de dados foi realizado com o auxílio do software Excel, onde foram registradas 1780 (instâncias) fichas cadastrais do serviço de calibração de balanças referentes ao ano de 2014 e 2015. A empresa estudada disponibilizou os documentos, os quais foram digitalizados e transcritos para a planilha. Neste caso, cada instância possui, em média, 60 variáveis, incluindo informações sobre a balança, condições ambientais e ensaios. Todas essas variáveis foram digitalizadas sem nenhuma seleção. Porém, a etapa de análise estatística dos dados não explorou todos esses dados e também demandou a elaboração de outras variáveis, tais como diferenças (variações) de temperatura, de pressão atmosférica e de umidade relativa do ar, além de codificações dos dados nominais para numéricos (APÊNDICE D).

Para os três problemas reais, na 1ª. fase (análise exploratória) foram realizados o teste T^2 de Hotelling, a estatística descritiva dos dados, a detecção da multicolinearidade dos dados, o descarte de atributos atípicos, Análise de Correlação e também Análise de Componentes Principais. Todavia, é indispensável que seja realizada uma análise dos dados para verificar se há falta de casos; se há presença de casos atípicos (*outliers*); se as hipóteses associadas à ferramenta escolhida estão sendo adequadamente atendidas, bem como identificar se os eventuais “afastamentos das condições ideais” poderão comprometer os resultados da análise.

Para a aplicação do teste T^2 de Hotelling foi utilizado as funções do software MATLAB. Através deste teste foi possível verificar que avaliação dos alunos perante ao curso em relação a “satisfeitos” e “insatisfeitos” são distintos, a um nível de significância de 5%, ou seja, uma probabilidade de 95% de acerto, pode-se dizer que os indivíduos referidos são distintos. Para o problema médico, também foi possível verificar que os pacientes com “câncer” ou “cálculo” no duto biliar, são

distintos. Para o problema de calibração da balança, foi possível verificar que a “boa qualidade” ou “baixa qualidade” são distintos.

Já para a estatística descritiva dos dados, detecção da multicolinearidade dos dados, e também o descarte de atributos atípicos foi utilizado o pacote estatístico SPSS 13.0. Com a estatística descritiva foi possível identificar os *outliers*, ou seja, que apresentavam características “fora do padrão”. Os *outliers* foram identificados pela análise de Escore Z ou distâncias de *Mahalanobis*. Esses *outliers* foram descartados, pois são considerados como atributos atípicos. Os pontos atípicos tiveram suas origens investigadas, e quando a área técnica ligada ao processo assumia que a natureza do ponto era externa ao processo, ele era eliminado.

Na 2ª. fase foram aplicadas as seguintes técnicas RL, FDLF, GSME-PL, RNAs e SVM para o DM, ou seja, para a classificação dos mesmos e, em seguida, foram obtidos e interpretados os seus resultados.

Para a 1ª. fase, 1ª base de dados (Lato Sensu), foi feito um teste com 885 instâncias, ou seja, não ocorreu a eliminação dos *outliers*. Para a 2ª base (diagnóstico médico) foram realizados dois testes: o primeiro, com a quantidade total de instâncias (118 instâncias da amostra) e o 2º. teste, após a eliminação dos *outliers,* com 97 instâncias. E finamente a 3ª base de dados (calibração de balança) foram realizados dois testes: o primeiro, com a quantidade total de instâncias (1540 instâncias da amostra) e o 2º teste, após a eliminação dos *outliers* com 1462 instâncias.

Na sequência, para a resolução do modelo de GSME-PL, utilizou-se o software LINGO (*Language Interactive General Optimizer*) e finalmente, para a aplicação da FDLF, RNAs e SVM foram utilizadas as funções do software *MATLAB*.

5 RESULTADOS

Os resultados propostos foram obtidos por meio da aplicação de duas fases para ambos os problemas reais analisados. A 1ª. fase, análise exploratória, ficou composta, do teste T^2 de Hotelling, Estatística Descritiva, Detecção de *outliers*, Análise de Correlação e também ACP. Já a 2ª. fase, aplicação das técnicas de DM, ficou composta, pelas técnicas: RL, GSME-PL, FDLF, RNAs e SVM.

5.1 INSTITUIÇÃO DE ENSINO SUPERIOR – PÓS GRADUAÇÃO LATO SENSU

5.1.1 Análise Exploratória de Dados

Para o primeiro problema real abordado, a análise exploratória dos dados foi aplicada, conforme já mencionado, com o intuito de “melhor entender” os atributos e “melhorar a qualidade” das instâncias obtendo-se, como consequência, a relevância (ou não) dos atributos, assim como uma maior acurácia da técnica de DM.

O teste T^2 de *Hotelling*, programado no *software MATLAB*, foi aplicado à amostra (885 instâncias e 12 atributos), com a obtenção dos seguintes valores: $\text{Amostra}_{(885\ \text{instâncias})}$: 83,7381 > 1,793 = $F_{12,872}$ (0,95). Por conseguinte, rejeita-se fortemente a hipótese de que as duas amostras (“satisfeitos” e “insatisfeitos”) estejam centradas no mesmo vetor de médias. Assim sendo, o conjunto de alunos “satisfeitos” com o curso é distinto do de alunos “insatisfeitos” com o curso.

Na sequência foi realizada a análise descritiva das 885 instâncias e, também, verificou se os atributos possuem correlação (ou não), por meio da aplicação do *software* SPSS 13.0. Os resultados encontram-se nas Tabelas 5.1 e 5.2, respectivamente, a seguir.

Na Tabela 5.1 tem-se na 1ª. coluna os nomes dos 12 atributos analisados; na 2ª. coluna está o número de instâncias válidas (885); na 3ª. coluna está a amplitude de cada um dos atributos; na 4ª., 5ª. e 6ª. colunas estão os valores mínimos, máximos e médios, respectivamente, de cada um dos atributos; na 7ª., 8ª. e 9ª. colunas estão os erros padrão, os desvios padrão e as variâncias, respectivamente, de cada um dos atributos.

Tabela 5.1 - Estatística descritiva das 885 Instâncias

Variáveis	Amplitude	Mínimo	Máximo	Média	Erro Padrão	Desvio padrão	Variância
Domínio do Conteúdo	3,250	6,750	10,000	9,432	0,016	**0,489**	0,239
Didática e Clareza na Condução do Módulo	4,500	5,500	10,000	9,180	0,022	**0,659**	0,434
Capacidade de despertar a motivação	6,208	4,375	10,583	8,927	0,028	**0,818**	0,669
Aderência do Conteúdo à Proposta do curso	4,500	5,500	10,000	9,209	0,020	**0,583**	0,340
Relacionamento do Professor com os Alunos	3,375	6,625	10,000	9,491	0,016	**0,475**	0,226
Planejamento e Organização geral	3,875	6,125	10,000	9,191	0,020	**0,603**	0,364
Sala de Aula	6,000	4,000	10,000	8,269	0,024	**0,720**	0,519
Eureka e intranet	5,750	3,750	9,500	7,589	0,028	**0,845**	0,713
Estrutura Cantinas e Banheiros	5,926	4,074	10,000	7,907	0,026	**0,761**	0,579
Tutor	5,143	4,857	10,000	8,071	0,023	**0,694**	0,482
Supervisão Acadêmica	5,143	4,857	10,000	8,180	0,021	**0,619**	0,383
Coordenação do Curso	5,846	4,154	10,000	8,356	0,022	**0,661**	0,438
Classe	1,000	0,000	1,000	0,364	0,016	**0,481**	0,232

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

A Tabela 5.1 destaca que os desvios atributos não estão acima de “3”, ou seja, não apontam a existência de dados atípicos (*outliers*), que deverão ser excluídos, assim sendo não será necessário aplicar os escores padronizados $z < -3$ ou $z > 3$ ou a distância *Mahalanobis* para o respectivo problema em questão.

Tabela 5.2 – Matriz de Correlação entre os 12 atributos

Variáveis		1	2	3	4	5	6	7	8	9	10	11	12
Correlação	1	1,000	0,875	0,799	0,807	0,795	0,810	0,147	0,045	0,089	0,157	0,186	0,183
	2	0,875	1,000	0,903	0,848	0,859	0,860	0,153	0,067	0,071	0,137	0,182	0,189
	3	0,799	0,903	1,000	0,819	0,846	0,809	0,140	0,054	0,083	0,167	0,209	0,210
	4	0,807	0,848	0,819	1,000	0,827	0,832	0,163	0,084	0,103	0,176	0,236	0,232
	5	0,795	0,859	0,846	0,827	1,000	0,822	0,167	0,055	0,088	0,162	0,202	0,209
	6	0,810	0,860	0,809	0,832	0,822	1,000	0,144	0,096	0,074	0,150	0,221	0,220
	7	0,147	0,153	0,140	0,163	0,167	0,144	1,000	0,373	0,606	0,350	0,352	0,357
	8	0,045	0,067	0,054	0,084	0,055	0,096	0,373	1,000	0,366	0,325	0,422	0,357
	9	0,089	0,071	0,083	0,103	0,088	0,074	0,606	0,366	1,000	0,372	0,380	0,351
	10	0,157	0,137	0,167	0,176	0,162	0,150	0,350	0,325	0,372	1,000	0,814	0,735
	11	0,186	0,182	0,209	0,236	0,202	0,221	0,352	0,422	0,380	0,814	1,000	0,888
	12	0,183	0,189	0,210	0,232	0,209	0,220	0,357	0,357	0,351	0,735	0,888	1,000

a. Variável dependente: Classe

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Na Tabela 5.2 observa-se que muitos dos atributos estão correlacionados. O atributo “domínio do conteúdo”, por exemplo, apresenta uma correlação de 0,875 em relação ao atributo “didática e clareza na condução do módulo” e assim tem-se que os seis primeiros atributos estão altamente correlacionados.

Desta forma, para que haja uma melhor interpretação dos dados, foi aplicada a ACP sobre os 12 atributos, cujos resultados estão apresentados na Tabela 5.3 a seguir. O 1º. Componente ficou composto pelos 6 primeiros atributos e poderia ser chamado, por exemplo, de “Docente”; o 2º. Componente reuniu os atributos 7, 8 e 9, podendo ser chamada de “Apoio Didático”; o 3º. Componente ficou formada pelos

atributos 10 e 12, “Infraestrutura” e finalmente, o 4º. Componente ficou constituída apenas do atributo “Eureka & internet”, aqui chamada de “Tecnologia e Informação – TI”.

Tabela 5.3 – Análise dos Componentes Principais

Variáveis	Componente			
	1	2	3	4
Domínio do Conteúdo	0,968			
Didática e Clareza na Condução do Módulo	0,928			
Capacidade de despertar a motivação	0,922			
Aderência do Conteúdo à Proposta do curso	0,922			
Relacionamento do Professor com os Alunos	0,914			
Planejamento e Organização Geral	0,914			
Coordenação do Curso		0,943		
Supervisão Acadêmica		0,932		
Tutor		0,918		
Estrutura Cantinas e Banheiros			0,895	
Sala de Aula			0,887	
Eureka e intranet				0,988
Método de extração: Análise do Componente principal.				
Método de rotação: Oblimin com normalização de Kaiser.[a]				
a. Rotação convergida em 4 iterações.				

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

A matriz de correlação para os 4 Componentes Principais está apresentada na Tabela 5.4 a seguir, que mostra que as componentes estão fracamente correlacionadas, ou seja, estão adequadas para a continuidade do processo.

Tabela 5.4 – Matriz de Correlação dos Componentes Principais

Componente	1	2	3	4
1	1,000	0,214	0,138	0,074
2	0,214	1,000	0,422	0,375
3	0,138	0,422	1,000	0,388
4	0,074	0,375	0,388	1,000
Método de extração: Análise do Componente principal.				
Método de rotação: Oblimin com normalização de Kaiser.				

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Uma interpretação gráfica para o problema pode ser visualizada nas Figuras 5.1 e 5.2 a seguir, que são autoexplicativas.

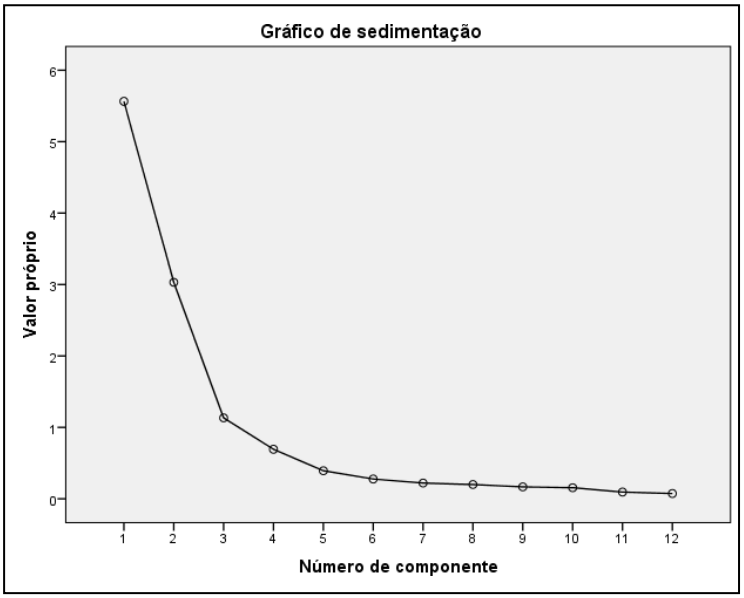

Figura 5.1 – Interpretação Gráfica para a obtenção dos 4 Componentes Principais
Fonte: Elaborado pela autora a partir do software SPSS 13.0

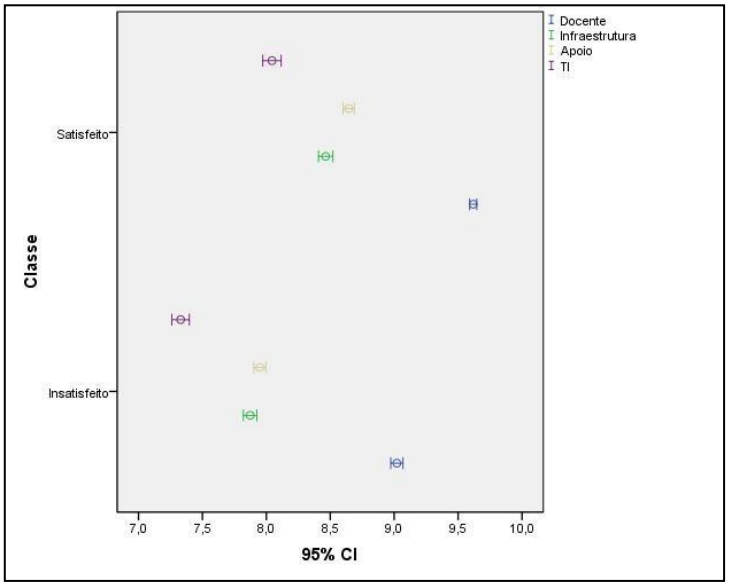

Figura 5.2 – Interpretação Gráfica para a obtenção dos 4 Componentes Principais
Fonte: Elaborado pela autora a partir do software SPSS 13.0

5.1.2 Regressão Logística Binária

A técnica de RLB foi aplicada, inicialmente, às 855 instâncias com 4 Componentes Principais, com o auxílio do *software* SPSS 13.0 utilizando o método "Entrada Forçada" (comando "*Enter*" no SPSS), que consiste na entrada simultânea de todos os componentes para definir o modelo final. O conjunto de dados foi dividido em duas partes: 70% dos dados para treinamento (619 instâncias), 30% para teste (266 instâncias).

A Tabela 5.5 apresenta os resultados iniciais da análise sem que qualquer variável independente do modelo tenha sido usada. As variáveis serão usadas mais adiante para que os resultados lá obtidos possam ser comparados com os da Tabela 5.5. No presente teste, todo aluno seria classificado como "insatisfeito" (porque houve um maior número de alunos respondendo como "insatisfeitos") e a taxa de acerto seria de 63,65%.

Tabela 5.5 – Classificação RLB (1º. Teste: 885 instâncias, sem variáveis)

Observado			Previsto		
			Classe		Porcentagem
			Insatisfeito	Satisfeito	
Etapa	Classe	Insatisfeito	394	0	100,0
		Satisfeito	225	0	0,0
	Porcentagem global				**63,65**

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

A Tabela 5.6 apresenta os testes de coeficientes do modelo *Omnibus* (ou, também chamado, de teste do ajustamento) que fornece uma indicação geral do desempenho do modelo com a inclusão das variáveis. Nesta Tabela 5.6 observa-se que todos os valores de *Sig*. estão em 0,000 (ou seja, $p < 0,0005$). Podemos concluir que o modelo com a inclusão das variáveis é melhor do que o anterior (Tabela 5.5). O valor de x^2 (chi-quadrado) é 557,201, com 4 graus de liberdade.

Tabela 5.6 – Verificação do ajuste do modelo RLB

		Qui-quadrado	df	Sig.
Etapa	Etapa	557,201	4,000	0,000
	Bloco	557,201	4,000	0,000
	Modelo	557,201	4,000	0,000

Fonte: Elaborado pela autora a partir do software SPSS 13.0.

Após 8 iterações na 1ª etapa, o modelo final selecionou os 4 Componentes Principais. A Tabela 5.7 mostra que no 1º. passo o índice "*R^2 de Cox e Snell*" situou-se no patamar de 59,3% e o "*R^2 Nagelkerke*" ficou em 81,3%. O "*R^2 de Cox e Snell"* indica que 59,3% das variações ocorridas na RLB são explicadas pelo conjunto dos 4 Componentes Principais, ou seja, este índice apresenta um alto índice de explicação. O índice *"R^2 Nagelkerke*" indica que 81,3% das variações registradas na variável dependente (Classe: "insatisfeitos" ou "satisfeitos") são ocasionadas pelos Componentes Principais. Ou seja, este índice também apresenta uma alta explicação. As magnitudes das duas estatísticas são consideráveis.

Tabela 5.7 – Testes para a verificação do ajuste do modelo RLB

Etapa	Verossimilhança de log -2	R quadrado Cox & Snell	R quadrado Nagelkerke
1	254,183[a]	0,593	0,813

Fonte: Elaborado pela autora a partir do software SPSS 13.0.

A Tabela 5.8 mostra o teste "*Hosmer e Lemeshow*" que também dá suporte ao modelo. Este teste, considerado por muitos como o mais confiável disponível para a avaliação do ajustamento do modelo, é interpretado de forma diferente do teste *omnibus* anterior. Para o teste de ajustamento *Hosmer-Lemeshow*, um ajustamento pobre é indicado por um valor *p* (ou *Sig*) < 0,05, ou seja, para que o ajustamento seja considerado adequado o valor *p* (ou *Sig*) ≥ 0,05. No nosso caso o valor *p* (ou *Sig*) = 0,491, ou seja, o modelo proposto está bem suportado.

Tabela 5.8 - Teste de Hosmer e Lemeshow

Etapa	Qui-quadrado	df	Sig.
1	7,433	8	0,491

Fonte: Elaborado pela autora a partir do software SPSS 13.0.

A Tabela 5.9 mostra os coeficientes B (2ª. coluna) que fazem a discriminação entre as duas classes. Assim, tem-se que η da equação (3.8) apresentada na seção (3.1.4), possui a forma mostrada em (5.1) para o problema apresentado.

$$fi_{(619)} = 6{,}947X_1 + 2{,}055X_2 + 3{,}575X_3 + 0{,}931X_4 - 119{,}646 \qquad (5.1)$$

onde: as variáveis X_i são os Componentes Principais (X_1 = Docente; X_2 = Infraestrutura; X_3 = Apoio; X_4= TI), sendo que todas elas são extremamente significativas, pois possuem o valor *p* (*Sig*) ≤ 0,05. Pela Tabela 5.3 já vista, considerando-se os 12 atributos do problema, tem-se que o relacionamento entre as variáveis originais e os componentes principais ocorre da seguinte forma: X_1 = [0,968 (atributo 1) + 0,928 (atributo 2) + 0,922 (atributo 3) + 0,922 (atributo 4) + 0,914 (atributo 5) + 0,914 (atributo 6)]; X_2 = [0,943 (atributo 7) + 0,932 (atributo 8) + 0,918 (atributo 9)]; X_3 = 0,895 [(atributo 10) + 0,887 (atributo 11)]; X_4 = 0,988 [(atributo 12)].

Tabela 5.9 – Coeficientes da RLB considerando as 619 instâncias e os 4 Componentes Principais

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Etapa 1ª	Docente	6,947	0,711	95,491	1,000	0,000	**1.040,098**	258,189	4.189,965
	Estrutura	2,055	0,327	39,437	1,000	0,000	7,807	4,111	14,827
	Apoio	3,575	0,454	61,932	1,000	0,000	35,709	14,657	86,995
	TI	0,931	0,249	14,007	1,000	0,000	2,537	1,558	4,132
	Constante	-119,646	10,997	118,374	1,000	0,000	0,000		

a. Variáveis inseridas na etapa 1: Docente, Estrutura, Apoio, TI.

Fonte: Elaborado pela autora a partir do software SPSS 13.0

Para cada valor apontado na 6ª coluna (*Exp(B)*) mostrado na Tabela 5.9 existe um intervalo de confiança de 95%, fornecendo um valor inferior e superior para *Exp (B)*, também conhecido como *odds ratio*. De acordo com Tabachnick e Fidell (2013), *odds ratio* representa “a chance de estar em uma das categorias quando o valor da variável preditora (independente) aumenta em uma unidade”. O valor *Exp (B)* é uma estimativa pontual do valor real, baseado em uma amostra, isto significa que quanto maior o Exp(B) maior a importância dos atributos. Como pode-se observar na Tabela 5.9, o preditor mais forte para o reporte é o Componente Principal “Docente”, com *Exp (B)* de 1.040,098, mostrando que a cada 1 aluno que avalie satisfatoriamente o componente “Docente”, aumentará em 1.040 as chances dele estar satisfeito com o curso. Da mesma forma, tem-se a interpretação para os demais valores da coluna de *Exp (B)*.

Já a matriz de classificação de treinamento para a RLB mostra uma taxa de acerto extremamente alta de instâncias classificadas corretamente para o modelo. Na Tabela 5.10, a taxa de acerto geral é de 91,0% para o treinamento e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “insatisfeitos”, de 92,1% e para a classe “satisfeitos”, de 88,9%. Assim, das 394 instâncias da classe “insatisfeitos”, apenas 31 foram erroneamente classificados como sendo da classe “satisfeitos” e das 225 instâncias consideradas como “satisfeitos”, o modelo classificou erroneamente 25 instâncias como “insatisfeitos”.

Tabela 5.10 - Treinamento de Classificação para as 619 instâncias e 4 Componentes Principais

Observado			Previsto		
			Classe		Porcentagem
			Insatisfeito	Satisfeito	
Etapa 1	Classe	Insatisfeito	**363**	31	92,1
		Satisfeito	25	**200**	88,9
	Porcentagem global				**91,0**

a. O valor de corte é ,500

Fonte: Elaborado pela autora a partir do software SPSS 13.0

No Quadro 5.1, a taxa de acerto geral é de 93,98% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “insatisfeitos”, de 90,53% e para a classe “satisfeitos”, de 100,0%. Assim, das 169 instâncias da classe “insatisfeitos”, apenas 16 foram erroneamente classificados como sendo da classe “satisfeitos” e das 97 instâncias consideradas como “satisfeitos”, o modelo não classificou erroneamente como “insatisfeitos”.

Quadro 5.1 – Matriz de confusão de teste com 266 instâncias para a RLB

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**153**	16	90,53%
Satisfeito	0	**97**	100,0%
		Porcentagem Global	**93,98%**

Fonte: Elaborado pela autora

5.1.3 Geração de uma Superfície que Minimiza Erros

Esta segunda técnica de DM constrói um modelo matemático que permite ajustar as variáveis do processo, no problema dos cursos de PG Latu Sensu, de forma a classificar alunos como “Satisfeito” ou “Insatisfeito” com um menor erro.

Foram classificados como alunos insatisfeitos, através da GSME-PL aqueles que forneceram um valor $A_w - e_m\,\gamma + y \geq e_m$, e alunos satisfeitos aqueles que forneceram um valor $-B_w + e_k\,\gamma + z \geq e_k$.

Deste modo, utilizou-se a mesma amostra (885 instâncias com 4 atributos), para construir um modelo que minimiza o custo, através das variáveis do processo, com a consequente determinação do hiperplano separador das classes: $w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \quad + w_nx_n = \gamma$.

O resultado obtido para a amostra mostra a função objetivo e a equação (5.2) que minimiza o erro:

Função Objetivo$_{(619)}$ = 1,742

Equação $_{(619)}$: $4{,}796X_1 + 1{,}492X_2 + 2{,}437X_3 + 0{,}744X_4 = 83{,}278$ (5.2)

A matriz de confusão para o treinamento e o teste para este caso está apresentada no quadro 5.2 e 5.3 a seguir.

Quadro 5.2 – Matriz de confusão de treinamento com 619 instâncias para a GSME-PL

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**354**	40	89,85%
Satisfeito	20	**205**	91,11%
		Porcentagem Global	**90,23%**

Fonte: Elaborado pela autora

Quadro 5.3 – Matriz de confusão de teste com 266 instâncias para a GSME-PL

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**143**	26	84,62%
Satisfeito	0	**97**	100,0%
		Porcentagem Global	**90,23%**

Fonte: Elaborado pela autora

No Quadro 5.2, a taxa de acerto geral é de 90,31% para o treinamento e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “insatisfeitos”, de 89,85% e para a classe “satisfeitos”, de 91,11%. No Quadro 5.3, a taxa de acerto geral é de 90,23% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “insatisfeitos”, de 84,62% e para a classe “satisfeitos”, de 100,0%. Assim, das 169 instâncias da classe “insatisfeitos”, apenas 26 foram erroneamente classificados como sendo da classe “satisfeitos” e das 97 instâncias consideradas como “satisfeitos”, o modelo não classificou erroneamente como “insatisfeitos”.

5.1.4 Função Discriminante Linear de Fisher

A FDLF possui função discriminante $Y = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ... + b_nX_n$, em que X_i, com $i = 1; ....;4$ representam cada uma das 4 características, e b_i, com $i = 1; ....; 4$, são os coeficientes ou pesos. Desta forma, para verificar se $x_0 \in A$ e se $x_0 \in B$, é necessário comparar o valor de Y com $Q = ½(x_A – x_B)' S_p^{-1}(x_A – x_B)$.

O resultado apresentado em (5.3) mostra a equação para a amostra com 619 instâncias com 4 atributos. A matriz de confusão de treinamento e teste para este caso estão apresentadas nos quadros 5.4 e 5.5 a seguir.

$Y_{(619)} = 3{,}560X_1 + 1{,}024X_2 + 2{,}147X_3 + 0{,}875X_4 < Q = 62{,}121$ e $Y = 3{,}560X_1 + 1{,}024X_2 + 2{,}147X_3 + 0{,}875X_4 > Q = 62{,}121$ (5.3)

Quadro 5.4 – Matriz de confusão de treinamento com 619 instâncias para a FDLF

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**342**	50	86,80%
Satisfeito	10	**215**	95,56%
		Porcentagem Global	**89,98%**

Fonte: Elaborado pela autora

Quadro 5.5 – Matriz de confusão de teste com 266 instâncias para a FDLF

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**48**	121	28,40%
Satisfeito	0	**97**	100,0%
		Porcentagem Global	**54,51%**

Fonte: Elaborado pela autora

Conforme o resultado dos quadros 5.4 e 5.5, têm-se que a amostra do conjunto de treinamento de A pertence a A somente 86,80% e a amostra do Conjunto B pertence a B somente 95,96% e a taxa de acerto geral é de 89,98%. A amostra do conjunto de teste de A pertence a A somente 28,40% e a amostra do Conjunto B pertence a B somente 100,0% e a taxa de acerto geral é de 54,51%, que não foi satisfatória comparado com a amostra de treinamento.

5.1.5 Redes Neurais Artificiais

Esta quarta técnica de DM fez uso do algoritmo de retropropagação do erro. Assim sendo, o treinamento é realizado em duas etapas. Na primeira etapa, um padrão ($m + k = 322 + 563 = 885$ instâncias) pertencente aos conjuntos A ou B deste trabalho. Para o problema dos cursos de PG Latu Sensu em relação a “satisfação” e “insatisfação” dos alunos, a rede neural, precisou de, aproximadamente, 100 iterações para convergir em cada uma das situações de teste. O conjunto de dados foi dividido em duas partes: 70% dos dados para treinamento e 30% para teste, usando 10 neurônios na camada oculta, pois foi o que gerou o melhor resultado entre 5, 10, 15, 20 e 25 na camada escondida.

Os resultados contidos na Figura 5.3 apresentam uma interpretação de treinamento de RNAs do *MATLAB*, onde observa-se as possibilidades de configuração de todos os parâmetros para as 619 instâncias (treinamento) e 4 atributos.

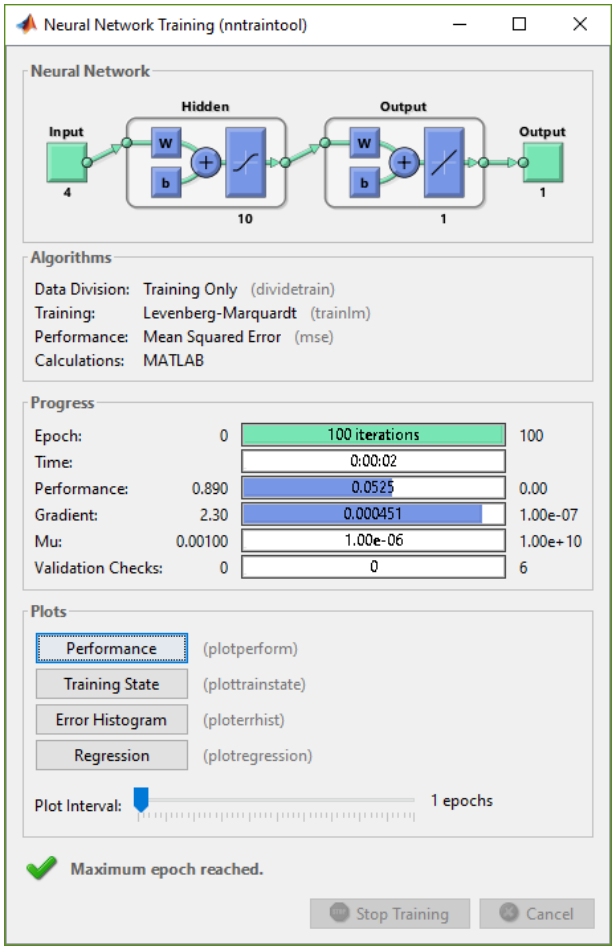

Figura 5.3 – Interpretação da RNAs com as 619 instâncias (PG Lato sensu)
Fonte: Elaborado pela autora a partir do *MATLAB Starter Application*

O desempenho da rede foi medido de acordo com a média dos quadrados dos erros (MSE) e erro absoluto médio (MAE).

Teste	Treinamento
$perf\_mse_{(619)} = 0{,}0525$	$perf\_mse_{(266)} = 0{,}0784$
$perf\_mae_{(619)} = 0{,}1349$	$perf\_mae_{(266)} = 0{,}1782$

A matriz de confusão de treinamento e teste para este caso está apresentada no quadro 5.6 e 5.7 a seguir.

Quadro 5.6 – Matriz de confusão de treinamento com 619 instâncias para a RNAs

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**368**	26	93,40%
Satisfeito	20	**205**	91,11%
		Porcentagem Global	**92,57%**

Fonte: Elaborado pela autora

Quadro 5.7 – Matriz de confusão de teste com 266 instâncias para a RNAs

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**148**	21	87,57%
Satisfeito	5	**92**	94,85%
		Porcentagem Global	**90,23%**

Fonte: Elaborado pela autora

5.1.6 Máquina de Vetor Suporte / *Support Vector Machine* (SVM)

Esta quinta técnica de DM é treinada com um algoritmo baseado na teoria estatística de aprendizagem, de forma a classificar os alunos como “Satisfeito” ou “Insatisfeito” com um menor erro. Deste modo, utilizou-se a equação do hiperplano que separa o conjunto de dados (conjunto A ou B) através da equação $w^t x + b = 0$, e assim poder definir as duas regiões que contêm cada uma das classes. Foi utilizado 70% dos dados para treinamento e 30% para teste.

O desempenho da rede foi medido de acordo com a média dos quadrados dos erros (MSE) e erro absoluto médio (MAE).

Teste	**Treinamento**
$perf\_mse_{(619)} = 0,0921$	$perf\_mse_{(266)} = 0,0639$
$perf\_mae_{(619)} = 0,0921$	$perf\_mae_{(266)} = 0,0639$

A matriz de confusão para este caso está apresentada no quadro 5.8 e 5.9 a seguir.

Quadro 5.8 – Treinamento de Classificação com 619 instâncias para a SVM

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**361**	33	91,62%
Satisfeito	24	**201**	89,33%
		Porcentagem Global	**90,79%**

Fonte: Elaborado pela autora

Quadro 5.9 – Matriz de confusão de teste com 266 instâncias para a SVM

	Insatisfeito	**Satisfeito**	**Porcentagem**
Insatisfeito	**153**	16	90,53%
Satisfeito	1	**96**	98,97%
		Porcentagem Global	**93,61%**

Fonte: Elaborado pela autora

5.1.7 Análise conjunta das técnicas

Desta forma, tem-se no Quadro 5.10 a seguir, o desempenho comparativo entre as cinco técnicas analisadas para o problema dos cursos de pós-graduação de uma IES privada. Observe-se que os 10 resultados de treinamento e teste para as 5 técnicas são bastante semelhantes, sendo que a diferença entre a melhor e pior acurácias é de 4,00%. De qualquer forma, para esta aplicação (PG Lato sensu), a RLB e VSM apresenta a maior acurácia (93,98% de acerto e 93,61%).

Quadro 5.10 – Comparação do desempenho das técnicas para o caso do problema dos cursos de PG de uma IES privada

Amostras	**RLB**		**GSME-PL**		**FDLF**		**RNAs**		**SVM**	
885	91,00%	**93,98%**	90,23%	90,31%	89,98%	54,51%	92,57%	90,23%	90,79%	**93,61%**

Fonte: Elaborado pela autora

5.2 PROBLEMA MÉDICO

5.2.1 Análise Exploratória de Dados

Para o segundo problema real abordado, problema médico, a análise exploratória dos dados foi aplicada, conforme já mencionado, com o intuito de filtrar os dados obtendo-se, como consequência, uma maior acurácia da técnica de DM. Para realização dos testes foram utilizadas duas amostras: a primeira com 118

instâncias (35 pertencentes à classe “câncer” e 83, à classe “cálculo”) e a segunda com 97 instâncias (28 da classe “câncer” e 69 da classe “cálculo”), cuja filtragem é mostrada mais adiante. A justificativa para se trabalhar com estas duas amostras (e não apenas uma) será mostrada mais adiante.

Incialmente, verificou-se se as variáveis independentes possuem correlação através da análise da multicolinariedade, por meio da aplicação do *software* SPSS 13.0. Os resultados encontram-se na Tabela 5.11, a seguir.

Tabela 5.11 - Coeficientes da Regressão

Variáveis	Coeficientes não padronizados		Coeficientes padronizados	t	Sig.	Estatísticas de colinearidade	
	B	Modelo padrão	Beta			Tolerância	VIF
(Constante)	0,233	0,472		0,494	0,623		
Idade	0,004	0,002	0,154	1,739	0,085	0,749	1,334
Sexo	0,150	0,082	0,161	1,826	0,071	0,755	1,324
Bilirrubina Direta	0,022	0,016	0,294	1,415	0,160	0,137	7,322
Bilirrubina Indireta	0,014	0,020	0,138	0,705	0,483	0,153	6,540
Fosfatases Alcalinas	0,000	0,001	-0,056	-0,406	0,686	0,310	3,222
SGOT	0,000	0,000	0,085	0,634	0,528	0,329	3,036
SGPT	0,000	0,000	0,024	0,235	0,815	0,544	1,839
Tempo de Atividade da Protrombin	0,000	0,000	-0,140	-1,766	0,080	0,940	1,064
Albumina	0,023	0,023	0,079	0,979	0,330	0,891	1,122
Amilase	0,023	0,067	0,029	0,337	0,737	0,776	1,288
Creatinina	-0,129	0,083	-0,129	-1,556	0,123	0,852	1,174
Leucócitos	-0,008	0,010	-0,070	-0,823	0,412	0,817	1,224
Volume Globular	-0,016	0,007	-0,208	-2,288	0,024	0,711	1,406

a. Variável dependente: DutoBiliar

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Nesta Tabela 5.11, têm-se na 1ª. coluna, a lista das variáveis; na 2ª. e 3ª. colunas, os coeficientes *Beta* (*B*) e os coeficientes padrão da Regressão; na 4ª. coluna, os coeficientes *B* padronizados; na 5ª. coluna, o teste *t* para cada variável; na 6ª. coluna, a Significância (Sig., ou ainda, valor *p*) e finalmente, na 7ª. e 8ª. colunas, a estatística de colinearidade, com os valores da Tolerância e do *Variance Inflation Factor* (VIF).

O SPSS identificou a variável bilirrubina total (atributo), apresentada na Tabela 5.11, como tendo forte correlação com as demais variáveis (Tolerância = 0) e, portanto, é recomendada a sua exclusão. Segundo Field (2009), quando o valor da Tolerância se aproxima de “0”, há forte indicação de multicolinearidade, o que diminui a qualidade do modelo. Vale ressaltar que o SPSS apresentou, automaticamente, duas tabelas, uma com as variáveis mantidas (Tabela 5.11) e

outra com as variáveis excluídas (neste caso, apenas uma: bilirrubina total; Tabela 5.12).

Tabela 5.12 - Variáveis excluídas

Variável	Beta In	t	Sig.	Correlação parcial	Estatísticas de colinearidade		
					Tolerância	VIF	Tolerância mínima
Bilirrubina Total	-9,829[b]	-1,159	0,249	-0,113	0,000	12.279,571	0,000

a. Variável dependente: DutoBiliar

b. Preditores no modelo: (Constante), Volume Globular, Tempo de Atividade da Protrombina, SGPT, Sexo, Albumina, Creatinina, SGOT, Amilase, Leucócitos, Bilirrubina Indireta, Idade, Fosfatases Alcalinas, Bilirrubina Direta

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Já a estatística descritiva, apresentada na Tabela 5.13 a seguir, foi realizada, inicialmente, para as 118 instâncias (35 pertencentes à classe "câncer" e 83 à classe "cálculo"). Esta Tabela 5.13 apresenta na 1ª. coluna a lista das variáveis; na 2ª. coluna, a quantidade de instâncias válidas (todos os 118); na 3ª. coluna, a amplitude dos valores de cada variável; na 4ª. e 5ª. colunas, são apresentados os valores mínimo e máximo assumidos por variável; na 6ª. coluna, o valor médio de cada variável; na 7ª. coluna, o erro padrão da média; nas colunas 8ª. e 9ª. colunas, os desvios padrões e as variâncias, respectivamente, para cada uma das variáveis.

Tabela 5.13 - Estatística descritiva das 118 instâncias

Variáveis	Amplitude	Mínimo	Máximo	Média	Erro padrão	**Desvio padrão**	Variância
Idade	66,000	17,000	83,000	50,915	1,544	**16,774**	281,360
Sexo	1,000	0,000	1,000	0,398	0,045	**0,492**	0,242
Bilirrubina Direta	28,200	0,300	28,500	7,472	0,562	**6,101**	37,219
Bilirrubina Indireta	20,200	0,400	20,600	5,041	0,411	**4,461**	19,896
Fosfatases Alcalinas	664,000	6,000	670,000	90,254	8,910	**96,786**	9.367,610
SGOT	1.210,000	10,000	1.220,000	95,203	11,636	**126,395**	15.975,633
SGPT	2.944,500	27,500	2.972,000	275,021	30,887	**335,514**	112.569,538
Tempo de Atividade da Protrombina	1.638,000	10,000	1.648,000	169,424	20,397	**221,565**	49.091,238
Albumina	12,000	11,000	23,000	14,034	0,148	**1,612**	2,597
Amilase	4,000	1,500	5,500	3,020	0,055	**0,598**	0,357
Creatinina	3,000	0,300	3,300	0,944	0,042	**0,459**	0,210
Leucócitos	24,300	3,700	28,000	9,497	0,359	**3,901**	15,215
Volume Globular	37,400	11,600	49,000	38,513	0,566	**6,143**	37,737
Duto Biliar	1,000	0,000	1,000	0,297	0,042	**0,459**	0,210

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Pode-se observar na tabela 5.13 que os desvios padrões para algumas variáveis estão acima de "3", ou seja, apontando a existência de dados atípicos

(*outliers*), os quais deverão ser excluídos, pois poderão influenciar negativamente e piorar o desempenho da técnica de DM. Assim, foram excluídos os dados que apresentaram os escores padronizados $z < -3$ ou $z > 3$ para cada uma das variáveis analisadas individualmente (esta técnica mostrou-se melhor que a distância *Mahalanobis* para o respectivo problema em questão). Foram excluídas 21 instâncias e, portanto, a amostra ficou com 97 instâncias (28 da classe "câncer" e 69 da classe "cálculo"). A Tabela 5.14 apresenta a estatística descritiva com os dados após a exclusão dos *outliers*. É possível notar, através da Tabela 5.14, que os desvios padrões das variáveis, analisados de forma conjunta, diminuíram após a exclusão dos 21 dados.

Tabela 5.14 - Estatística descritiva dos dados das 97 instâncias (após a exclusão dos *outliers*)

Variáveis	Amplitude	Mínimo	Máximo	Média	Erro padrão	**Desvio padrão**	Variância
Idade	66,000	17,000	83,000	52,629	1,727	**17,005**	289,173
Sexo	1,000	0,000	1,000	0,381	0,050	**0,488**	0,238
Bilirrubina Direta	19,080	0,300	19,380	6,818	0,525	**5,169**	26,722
Bilirrubina Indireta	15,600	0,400	16,000	4,700	0,410	**4,043**	16,345
Fosfatases Alcalinas	294,000	6,000	300,000	85,000	7,926	**78,060**	6.093,292
SGOT	360,000	10,000	370,000	83,041	6,694	**65,929**	4.346,623
SGPT	1.020,850	29,150	1.050,000	235,284	18,381	**181,028**	32.771,277
Tempo de Atividade da protrombina	796,000	18,000	814,000	138,412	11,611	**114,359**	13.078,078
Albumina	6,000	12,000	18,000	13,856	0,121	**1,190**	1,416
Amilase	2,800	1,700	4,500	3,019	0,055	**0,541**	0,293
Creatinina	1,600	0,400	2,000	0,880	0,029	**0,283**	0,080
Leucócitos	16,800	3,700	20,500	9,103	0,309	**3,044**	9,268
Volume Globular	20,400	28,600	49,000	39,514	0,432	**4,255**	18,102
Duto Biliar	1,000	0,000	1,000	0,289	0,046	**0,455**	0,207

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Na sequência, o teste T^2 de Hotelling foi aplicado às duas amostras (118 e 97 instâncias), com a obtenção dos seguintes valores: $Amostra_{(118\ instâncias)}$: 5,09 > 1,819 = $F_{13,104}$ (0,95); $Amostra_{(97\ instâncias)}$: 6,32 > 1,845 = $F_{13,83}$ (0,95). Por conseguinte, rejeita-se fortemente para as duas amostras, a hipótese de que as mesmas estejam centradas no mesmo vetor de médias. Assim sendo, o conjunto de pacientes com "câncer" no duto biliar é distinto do de "cálculo" no duto biliar.

Têm-se, assim, duas amostras para se aplicar a RLB: a primeira, com os dados originais, possui 118 instâncias (35 pertencentes à classe "câncer" e 83, à classe "cálculo") e a segunda, com os dados analisados estatisticamente (dados transformados), possui 97 instâncias (28 da classe "câncer" e 69 da classe "cálculo").

5.2.2 Regressão Logística Binária

A técnica de DM analisada RLB, foi aplicada por meio da realização de um teste. Para o teste foi utilizada, primeiramente, a amostra com 97 instâncias. O 1º. teste se mostrou satisfatório, ou seja, “suficiente” para obter um desempenho aceitável da RL, devido a quantidade de atributos.

A técnica de DM, RLB, foi aplicada, inicialmente, às 118 instâncias, com o auxílio do *software* SPSS 13.0 utilizando o método “Entrada Forçada” (comando “*Enter*” no SPSS), que consiste na entrada simultânea de todas as variáveis para definir o modelo final que minimiza o número de variáveis e maximiza a precisão do modelo.

A matriz de confusão (Tabela 5.15) apresenta a classificação para as 83 instâncias. A taxa de acerto global foi de 80,7% e as taxas individuais de acertos foram: para a classe “cálculo”, de 86,2% e para a classe “câncer”, de 68,0%. Assim, dos 58 padrões da classe cálculo, apenas 8 estavam sendo classificados como sendo “câncer” e dos 25 padrões considerados câncer, 8 estavam sendo classificados como “cálculo”.

Tabela 5.15 - RLB: Treinamento de Classificação para as 83 instâncias

Observado			Previsto		
			Duto Biliar		Porcentagem
			Cálculo	Cancêr	
Etapa 1	Duto Biliar	Cálculo	**50**	8	86,2
		Cancêr	8	**17**	68,0
	Porcentagem global				**80,7**

a. O valor de corte é ,500

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

No Quadro 5.11, a taxa de acerto geral é de 82,86% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 76,0% e para a classe “câncer”, de 100,0%. Assim, das 25 instâncias da classe “cálculo”, apenas 6 foram erroneamente classificados como sendo da classe “câncer” e das 10 instâncias consideradas como “câncer”, o modelo não classificou erroneamente como “cálculo”.

Quadro 5.11 – Matriz de confusão de teste com 35 instâncias para a RLB

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**19**	6	76,0%
Câncer	0	**10**	100,0%
Porcentagem Global			**82,86%**

Fonte: Elaborado pela autora

A Tabela 5.16 mostra os coeficientes B (2ª. coluna) que fazem a discriminação entre as classes. Assim, têm-se que η da equação (3.8) apresentada na seção 3.1.4, possui a forma mostrada em (5.4), a seguir, onde as variáveis X_i são as variáveis do problema (X_1 = idade; ...; X_{13} = volume globular).

Tabela 5.16 - Coeficientes da RLB considerando os 83 instâncias e 13 atributos (desconsiderada a "bilirrubina total")

	Variáveis	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Etapa 1[a]	Idade	-0,016	0,030	0,286	1,000	0,593	0,984	0,929	1,043
	Sexo	2,063	1,087	3,602	1,000	0,058	7,868	0,935	66,231
	Bilirrubina Direta	1,198	0,497	5,809	1,000	0,016	3,312	1,251	8,770
	Bilirrubina Indireta	-0,952	0,482	3,912	1,000	0,048	0,386	0,150	0,991
	Fosfatases Alcalinas	-0,011	0,008	1,815	1,000	0,178	0,989	0,972	1,005
	SGOT	0,007	0,006	1,633	1,000	0,201	1,007	0,996	1,018
	SGPT	0,011	0,004	7,980	1,000	0,005	1,011	1,003	1,018
	Tempo de Atividade da Protrombina	-0,013	0,006	4,413	1,000	0,036	0,987	0,975	0,999
	Albumina	0,092	0,196	0,220	1,000	0,639	1,097	0,746	1,612
	Amilase	-1,592	1,108	2,064	1,000	0,151	0,204	0,023	1,785
	Creatinina	-3,883	1,983	3,832	1,000	0,050	0,021	0,000	1,005
	Leucócitos	-0,015	0,139	0,011	1,000	0,915	0,985	0,751	1,293
	Volume Globular	-0,181	0,089	4,121	1,000	0,042	0,834	0,701	0,994
	Constante	8,613	5,461	2,488	1,000	0,115	5.501,736		

a. Variáveis inseridas na etapa 1: Idade, Sexo, Bilirrubina direta, Bilirrubina indireta, Fosfatases alcalinas, SGOT, SGPT, Tempo de atividade da protrombina, Albumina, Amilase, Creatinina, Leucócitos, Volume globular.

Fonte: Elaborado pela autora a partir do software SPSS 13.0

$$fi_{(83)} = - 0{,}016X_1 + 2{,}063X_2 + 1{,}198X_3 - 0{,}952X_4 - 0{,}011X_5 + 0{,}007X_6 + 0{,}011X_7 - 0{,}013X_8 + 0{,}092X_9 - 1{,}592X_{10} - 3{,}883X_{11} - 0{,}015X_{12} - 0{,}181X_{13} + 8{,}613 \quad (5.4)$$

Um novo teste foi aplicado na amostra com 97 instâncias. A Tabela 5.17 de treinamento apresenta o resultado inicial considerando o modelo com apenas uma constante, ou seja, se todo paciente fosse classificado como "cálculo", a taxa de acerto seria de 70,6%, considerado insatisfatório.

Tabela 5.17 - Classificação segundo teste treinamento (68 instâncias e 13 atributos)

Observado			Previsto		
			DutoBiliar		Porcentagem
			Cálculo	Câncer	
Etapa 0	DutoBiliar	Cálculo	48	0	100,0
		Câncer	20	0	0,0
	Porcentagem global				**70,6**

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

O teste *Omnibus* de coeficientes do modelo (ou, também chamado, de teste de ajustamento), conforme a Tabela 5.18, fornece uma indicação geral do desempenho do modelo. Observa-se que todos os valores de Sig. (Significância) estão em 0,000 (ou seja, $p < 0,0005$). Podemos concluir que o modelo com a exclusão dos 21 dados pode ser utilizado. O valor de X^2 (chi-quadrado) é 58,178, com 13 graus de liberdade.

Tabela 5.18 - Testes de coeficientes de modelo *Omnibus*

		Qui-quadrado	df	Sig.
Etapa 1	Etapa	58,178	13	0,000
	Bloco	58,178	13	0,000
	Modelo	58,178	13	0,000

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Após 9 iterações na 1ª etapa, o modelo final selecionou todas as 13 variáveis. A Tabela 5.19 mostra que na 1ª etapa o índice "R^2 *Cox* e *Snell*" situou-se no patamar de 57,5% e o "R^2 *Nagelkerke*" ficou em 81,9%. O "R^2 *Cox* e *Snell*" indica que 57,5% das variações ocorridas na variável dependente ("câncer" ou "cálculo" no duto biliar) são explicadas pelo conjunto das variáveis independentes, ou seja, este índice apresenta um alto índice de explicação. Da mesma forma, o índice "R^2 *Nagelkerke*" indica que 81,9% das variações registradas na variável dependente são explicadas pelas variáveis independentes. Ou seja, este índice também apresenta uma alta explicação.

Tabela 5.19 - Resumo do modelo

Etapa	Verossimilhança de log -2	R quadrado Cox & Snell	R quadrado Nagelkerke
1	24,210[a]	0,575	0,819

Fonte: Elaborado pela autora a partir do software SPSS 13.0

A Tabela 5.20 mostra que o teste "*Hosmer e Lemeshow*" também dá suporte ao modelo. Este teste, considerado por muitos como o mais confiável disponível para a avaliação do ajustamento do modelo, é interpretado de forma diferente do

teste *Omnibus* anterior. Para o teste de ajustamento *Hosmer-Lemeshow*, um ajustamento pobre é indicado por um valor *p* (ou Sig) < 0,05, ou seja, para que o ajustamento seja considerado adequado o valor *p* (ou Sig) ≥ 0,05. No nosso caso o valor *p* (ou Sig) = 0,924, ou seja, o modelo proposto está bem suportado.

Tabela 5.20 – Teste de Hosmer e Lemeshow

Etapa	Qui-quadrado	df	Sig.
1	3,161	8	0,924

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

A Tabela 5.21 mostra os coeficientes B (2ª. coluna) para a equação (3.8) que modelará a discriminação entre as classes. Em (5.5), a seguir, têm-se a expressão de η da equação (3.8) apresentada na seção 3.1.4.

$$fi_{(68)} = - 0{,}053X_1 + 3{,}911X_2 + 2{,}336X_3 – 1{,}749X_4 - 0{,}023X_5 - 0{,}007X_6 + 0{,}014X_7 - 0{,}041X_8 + 2{,}534X_9 - 1{,}517X_{10} + 0{,}716X_{11} - 0{,}222X_{12} - 0{,}710X_{13} – 5{,}144 \quad (5.5)$$

onde as variáveis X_i são as apresentadas na Tabela 5.21 (X_1 = idade; ...; X_{13} = volume globular), a seguir. Além disso, esta Tabela 5.21 mostra que, na verdade, apenas as variáveis “bilirrubina direta” e “volume globular” são estatisticamente significativas, fato que pode ser constatado por meio da coluna “Sig.”). Variáveis com valores inferiores a 0,05 nesta coluna são significativas.

Tabela 5.21 – Coeficientes da RLB considerando os 68 instâncias e 13 atributos (desconsiderada a “bilirrubina total”)

	Variáveis	B	S.E.	Wald	df	Sig.	Exp(B)	EXP(B)	
								Inferior	Superior
Etapa 1ª	Idade	-0,053	0,064	0,686	1,000	0,408	0,948	0,837	1,075
	Sexo	3,911	2,553	2,347	1,000	0,125	49,953	0,335	7.438,115
	Bilirrubina Direta	2,336	1,211	3,724	1,000	0,054	**10,341**	0,964	110,925
	Bilirrubina Indireta	-1,749	1,044	2,808	1,000	0,094	0,174	0,022	1,345
	Fosfatases Alcalinas	-0,023	0,015	2,417	1,000	0,120	0,977	0,950	1,006
	SGOT	-0,007	0,020	0,135	1,000	0,713	0,993	0,955	1,032
	SGPT	0,014	0,008	3,482	1,000	0,062	1,014	0,999	1,029
	Tempo de atividade da Protrombina	-0,041	0,024	2,936	1,000	0,087	0,960	0,916	1,006
	Albumina	2,534	1,753	2,090	1,000	0,148	12,603	0,406	391,310
	Amilase	-1,517	2,288	0,439	1,000	0,507	**0,219**	0,002	19,453
	Creatinina	0,716	3,120	0,053	1,000	0,818	2,046	0,005	925,458
	Leucócitos	-0,222	0,300	0,545	1,000	0,460	0,801	0,445	1,443
	Volume Globular	-0,710	0,348	4,153	1,000	0,042	0,492	0,249	0,973
	Constante	-5,144	18,350	0,079	1,000	0,779	0,006		

a. Variáveis inseridas na etapa 1: Idade, Sexo, Bilirrubina direta, Bilirrubina indireta, Fosfatases alcalinas, SGOT, SGPT, Tempo de atividade da protrombina, Albumina, Amilase, Creatinina, Leucócitos, Volume globular.

Fonte: Elaborado pela autora a partir do software SPSS 13.0

Para cada valor apontado na 6ª coluna (Exp(B)) mostrado na Tabela 5.21 existe um intervalo de confiança de 95%, fornecendo um valor inferior e superior para Exp (B), também conhecido como *odds ratio*. De acordo com Tabachnick e Fidell (2013), *odds ratio* representa “a chance de estar em uma das categorias quando o valor da variável preditora (independente) aumenta em uma unidade”. O valor Exp (B) é uma estimativa pontual do valor real, baseado em uma amostra.

Considerando-se as variáveis com significância estatística (“bilirrubina direta” e “amilase”) tem-se, pela Tabela 5.21, que a cada unidade a mais de “bilirrubina direta”, aumentará em 10,341 vezes a probabilidade do diagnóstico do paciente ser câncer e a cada unidade a mais de “amilase”, diminuirá em 0,219 vezes a probabilidade do diagnóstico do paciente ser câncer. Vale enfatizar que neste caso, pelo fato do valor (0,219) ser menor do que “1”, a interpretação é a inversa da anterior. Na última coluna desta Tabela 5.21, tem-se o intervalo de confiança para estes valores: [0,964; 110,925] e [0,002; 19,453] para a bilirrubina direta e para a amilase, respectivamente.

As matrizes de classificação mostram uma taxa de acurácia satisfatória. Na Tabela 5.22, a taxa de acerto geral é de 94,1% e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 97,9% e para a classe “câncer”, de 85%. Assim, das 48 instâncias da classe “cálculo”, apenas 1 estão na classificação de “câncer" e das 20 instâncias considerados “câncer” detinham 3 instâncias “cálculo”.

Tabela 5.22 - RLB: Treinamento de Classificação para as 68 instâncias

Observado			Previsto		
			Duto Biliar		Porcentagem
			Cálculo	Câncer	
Etapa 1	Duto Biliar	Cálculo	**47**	1	97,9
		Câncer	3	**17**	85,0
	Porcentagem global				**94,1**

Fonte: Elaborado pela autora a partir do software SPSS 13.0

No Quadro 5.12, a taxa de acerto geral é de 79,31% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 80,95% e para a classe “câncer”, de 75,0%. Assim, das 21 instâncias da classe “cálculo”, apenas 4 foram erroneamente classificados como sendo da classe

“câncer” e das 8 instâncias consideradas como “câncer”, o modelo classificou 2 erroneamente como “cálculo”.

Quadro 5.12 – Matriz de confusão de teste com 29 instâncias para a RLB

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**17**	4	80,95%
Câncer	2	**6**	75,0%
	Porcentagem Global		**79,31%**

Fonte: Elaborado pela autora

5.2.3 Geração de uma Superfície que Minimiza Erros

Esta segunda técnica de DM constrói um modelo matemático que permite ajustar as variáveis do processo, no problema médico, de forma a classificar pacientes com “câncer” ou “cálculo” no duto biliar com um menor erro.

Foram classificados os pacientes com cálculo no duto biliar, através da GSME-PL aqueles que forneceram de um valor $A_w - e_m\,\gamma + y \geq e_m$, e pacientes com câncer no duto biliar aqueles que forneceram um valor $- B_w + e_k\,\gamma + z \geq e_k$.

Deste modo, utilizou-se as mesmas duas amostras (118 e 97 instâncias), para construir um modelo que minimiza o custo, através das variáveis do processo, com a consequente determinação do hiperplano separador das classes: $w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \quad + w_nx_n = \gamma$.

Os resultados apresentados em (5.6) e (5.7) mostram os valores da função objetivo e, também as equações para as amostras com 83 e 68 instâncias respectivamente. As matrizes de confusão para as duas instâncias estão apresentadas nos Quadros 5.13 e 5.16, a seguir.

Função Objetivo$_{(83)}$ = 0,621

Equação$_{(83)}$: $0{,}007X_1 + 2{,}037X_2 + 0{,}917X_3 - 0{,}737X_4 - 0{,}008X_5 + 0{,}004X_6 + 0{,}007X_7 - 0{,}014X_8 + 0{,}171X_9 - 0{,}854X_{10} - 2{,}867X_{11} + 0{,}020X_{12} - 0{,}161X_{13} = - 5{,}455$ (5.6)

A matriz de confusão para este caso está apresentada no quadro 5.13 e 5.14 a seguir.

Quadro 5.13 – Treinamento de Classificação com 83 instâncias para a GSME-PL

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**46**	12	79,31%
Câncer	1	**24**	96,0%
Porcentagem Global			**84,34%**

Fonte: Elaborado pela autora

Quadro 5.14 – Matriz de confusão de teste com 35 instâncias para a GSME-PL

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**16**	6	76,0%
Câncer	1	**9**	90,0%
Porcentagem Global			**80,0%**

Fonte: Elaborado pela autora

No Quadro 5.13, a taxa de acerto geral é de 84,34% para o treinamento e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 79,31% e para a classe “câncer”, de 96,0%. No Quadro 5.14, a taxa de acerto geral é de 80,0% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 76,0% e para a classe “câncer”, de 90,0%.

Função Objetivo$_{(68)}$ = 0,400

Equação minimiza erro$_{(68)}$: $- 0{,}021X_1 + 1{,}461X_2 + 0{,}814X_3 - 0{,}494X_4 - 0{,}013X_5 - 0{,}011X_6 + 0{,}009X_7 - 0{,}022X_8 + 1{,}547X_9 - 1{,}008X_{10} + 1{,}174X_{11} - 0{,}266X_{12} - 0{,}311X_{13} = 5{,}010$ (5.7)

A matriz de confusão para este caso está apresentada no quadro 5.15 e 5.16 a seguir.

Quadro 5.15 – Treinamento de Classificação com 68 instâncias para a GSME-PL

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**42**	6	87,50%
Câncer	2	**18**	90,0%
Porcentagem Global			**88,24%**

Fonte: Elaborado pela autora

Quadro 5.16 – Matriz de confusão de teste com 29 instâncias para a GSME-PL

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**15**	6	71,43%
Câncer	2	**6**	75,0%
Porcentagem Global			**72,41%**

Fonte: Elaborado pela autora

No Quadro 5.15, a taxa de acerto geral é de 88,24% para o treinamento e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 87,50% e para a classe “câncer”, de 90,0%. No Quadro 5.16, a taxa de acerto geral é de 72,41% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “cálculo”, de 71,43% e para a classe “câncer”, de 75,0%.

Conforme o resultado tem-se que as amostras de treinamento com 68 instâncias (88,24%) mostraram um desempenho um pouco melhor em relação à amostra de treinamento com 83 instâncias (84,34%).

5.2.4 Função Discriminante Linear de Fisher

A FDLF possui função discriminante $Y = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ... + b_nX_n$, em que X_i, com $i = 1; ... ; 13$ representa cada uma das 13 variáveis e b_i, com $i = 1; ... ; 13$ são os coeficientes ou pesos. Desta forma, para verificar se $x_0 \in A$ ou *se* $x_0 \in B$, é necessário comparar o valor de Y com $q = ½(x_A – x_B)' S_p^{-1}(x_A – x_B)$.

Os resultados apresentados em (5.8) e (5.9) mostram as equações para as amostras com 83 e 68 instâncias, respectivamente. As matrizes de confusão para as duas instâncias estão apresentadas nos Quadros 5.17 e 5.20, a seguir.

$$Y_{(83)} = 0{,}012X_1 + 1{,}359X_2 + 0{,}563X_3 - 0{,}325X_4 - 0{,}008X_5 + 0{,}004X_6 + 0{,}0009X_7 - 0{,}002X_8 + 0{,}158X_9 - 0{,}219X_{10} – 1{,}227X_{11} – 0{,}109X_{12} – 0{,}152X_{13} < q = - 1{,}1215 \text{ e}$$

$$Y_{(83)} = 0{,}012X_1 + 1{,}359X_2 + 0{,}563X_3 - 0{,}325X_4 - 0{,}008X_5 + 0{,}004X_6 + 0{,}0009X_7 - 0{,}002X_8 + 0{,}158X_9 - 0{,}219X_{10} – 1{,}227X_{11} – 0{,}109X_{12} – 0{,}152X_{13} \geq q = - 1{,}1215 \quad (5.8)$$

A matriz de confusão para este caso está apresentada no quadro 5.17 e 5.18 a seguir.

Quadro 5.17 – Treinamento de Classificação com 83 instâncias para a FDLF

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**48**	10	82,76%
Câncer	4	**21**	84,0%
Porcentagem Global			**83,13%**

Fonte: Elaborado pela autora

No Quadro 5.17, têm-se que a amostra do conjunto A pertence a A somente 82,76% e a amostra do Conjunto B pertence a B somente 84,00%.

Quadro 5.18 – Matriz de confusão de teste com 35 instâncias para a FDLF

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**17**	8	68,0%
Câncer	1	**9**	90,0%
Porcentagem Global			**74,29%**

Fonte: Elaborado pela autora

No Quadro 5.18, têm-se que a amostra do conjunto A pertence a A somente 68,0% e a amostra do Conjunto B pertence a B somente 90,0%.

$$Y_{(68)} = -0{,}013X_1 + 1{,}151X_2 + 0{,}843X_3 - 0{,}433X_4 - 0{,}017X_5 + 0{,}007X_6 + 0{,}0008X_7 - 0{,}012X_8 + 1{,}188X_9 - 0{,}720X_{10} - 0{,}404X_{11} - 0{,}035X_{12} - 0{,}336X_{13} < q = 3{,}921 \text{ e}$$

$$Y_{(68)} = -0{,}013X_1 + 1{,}151X_2 + 0{,}843X_3 - 0{,}433X_4 - 0{,}017X_5 + 0{,}007X_6 + 0{,}0008X_7 - 0{,}012X_8 + 1{,}188X_9 - 0{,}720X_{10} - 0{,}404X_{11} - 0{,}035X_{12} - 0{,}336X_{13} \geq q = -3{,}921 \qquad (5.9)$$

A matriz de confusão para este caso está apresentada no quadro 5.19 e 5.20 a seguir.

Quadro 5.19 – Treinamento de Classificação com 68 instâncias para a FDLF

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**40**	8	83,33%
Câncer	2	**18**	90,0%
Porcentagem Global			**85,29%**

Fonte: Elaborado pela autora

No Quadro 5.19, têm-se que a amostra do conjunto A pertence a A somente 83,33% e a amostra do Conjunto B pertence a B somente 90,0%.

Quadro 5.20 – Matriz de confusão de teste com 35 instâncias para a FDLF

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**16**	5	76,19%
Câncer	2	**6**	75,0%
Porcentagem Global			**75,86%**

Fonte: Elaborado pela autora

No Quadro 5.20, têm-se que a amostra do conjunto A pertence a A somente 76,19% e a amostra do Conjunto B pertence a B somente 75,0%.

Conforme o resultado tem-se que a amostra com 68 instâncias de treinamento (85,29%) mostrou um desempenho um pouco melhor em relação à amostra com 83 instâncias de treinamento (83,13%).

5.2.5 Redes Neurais Artificiais

Esta quarta técnica de DM fez uso do algoritmo de retropropagação do erro. Assim sendo, o treinamento é realizado em duas etapas. Na primeira etapa, um padrão ($m + k = 83 + 35 = 118$ instâncias) e ($m + k = 69 + 28 = 97$ instâncias) pertencente aos conjuntos A ou B deste trabalho. Para classificar os pacientes com “Cálculo” ou “Câncer” com um menor erro, precisou de 18 iterações para 83 instâncias e 13 iterações para 68 instâncias, e assim convergir em cada uma das situações de teste (sabendo-se que o critério de convergência foi: $1{,}865e^{-23}$ para 83 instâncias (treinamento) e $1{,}550e^{-26}$ para 68 instâncias (teste)). O conjunto de dados foi dividido em duas partes: 70% dos dados para treinamento e 30% para teste, usando 10 neurônios na camada oculta, pois foi o que gerou o melhor resultado entre 5, 10, 15, 20 e 25 na camada escondida.

Os resultados, apresentados nas Figuras 5.4 e 5.5, apresentam uma interpretação do treinamento das RNAs, executadas no *MATLAB*, onde observa-se as possibilidades de configuração de todos os parâmetros para as 83 instâncias e 68 instâncias e com 13 atributos ambas.

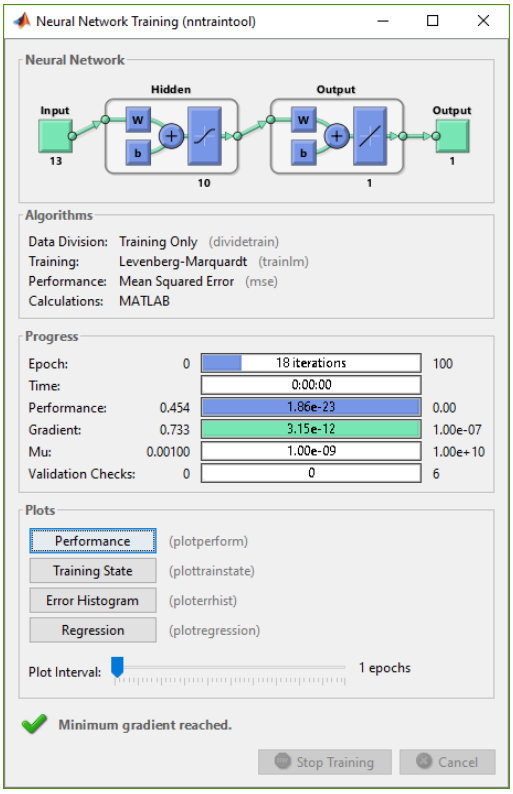

Figura 5.4 – Interpretação da RNAs com as 83 instâncias
Fonte: Elaborado pela autora a partir do *MATLAB Starter Application*

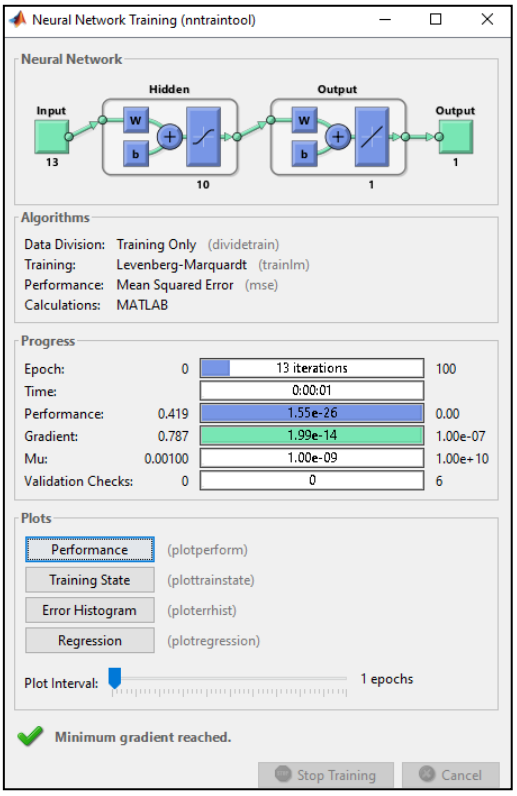

Figura 5.5 – Interpretação da RNAs com as 68 instâncias
Fonte: Elaborado pela autora a partir do *MATLAB Starter Application*

O desempenho da rede foi medido de acordo com a média dos quadrados dos erros (MSE) e erro absoluto médio (MAE).

Teste	Treinamento
perf_mse(83) = $1{,}865e^{-23}$	perf_mse(35) = 1,662
perf_mae(83) = $3{,}215e^{-12}$	perf_mae(35) = 0,890

Teste	Treinamento
perf_mse(35) = 1,662	perf_mse(29) = 0,327
perf_mae(35) = 0,890	perf_mae(29) = 0,444

A matriz de confusão de treinamento e teste para estes dois casos estão apresentadas nos quadros 5.21 e 5.24 a seguir.

Quadro 5.21 – Treinamento de Classificação com 83 instâncias para a RNAs

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**58**	0	100,0%
Câncer	0	**25**	100,0%
Porcentagem Global			**100,0%**

Fonte: Elaborado pela autora

Quadro 5.22 – Matriz de confusão de teste com 35 instâncias para a RNAs

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**16**	9	64,0%
Câncer	2	**8**	80,0%
Porcentagem Global			**68,57%**

Fonte: Elaborado pela autora

Quadro 5.23 – Treinamento de Classificação com 68 instâncias para a RNAs

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**48**	0	100,0%
Câncer	0	**20**	100,0%
Porcentagem Global			**100,0%**

Fonte: Elaborado pela autora

Quadro 5.24 – Matriz de confusão de teste com 29 instâncias para a RNAs

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**16**	5	76,19%
Câncer	3	**5**	62,50%
Porcentagem Global			**72,41%**

Fonte: Elaborado pela autora

Conforme o resultado tem-se que a amostra com 83 instâncias treinamento mostrou o mesmo desempenho em relação à amostra com 68 instâncias treinamento (100,0%).

5.2.6 Máquina de Vetor Suporte / Support Vector Machine (SVM)

Esta quinta técnica de DM é treinada com um algoritmo baseado na teoria estatística de aprendizagem, de forma a classificar os pacientes com “Cálculo” ou “Câncer” com um menor erro. Deste modo, utilizou-se a equação do hiperplano que separa o conjunto de dados (conjunto A ou B) através da equação $w^T x + b = 0$, e assim poder definir as duas regiões que contêm cada uma das classes, foi utilizado 70% para treinamento e 30% para teste.

O desempenho da rede foi medido de acordo com a média dos quadrados dos erros (MSE) e erro absoluto médio (MAE).

Teste	**Treinamento**
perf_mse(83) = 0,169	perf_mse(35) = 0,200
perf_mae(83) = 0,169	perf_mae(35) = 0,200

Teste	**Treinamento**
perf_mse(68) = 0,044	perf_mse(29) = 0,172
perf_mae(68) = 0,044	perf_mae(29) = 0,172

A matriz de confusão de treinamento e teste para estes dois casos estão apresentadas nos quadros 5.25 e 5.28 a seguir.

Quadro 5.22 – Treinamento de Classificação com 83 instâncias para a VSM

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**53**	5	91,38%
Câncer	9	**16**	64,0%
Porcentagem Global			**83,13%**

Fonte: Elaborado pela autora

Quadro 5.23 – Matriz de confusão de teste com 35 instâncias para a VSM

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**19**	6	76,0%
Câncer	1	**9**	90,0%
Porcentagem Global			**80,0%**

Fonte: Elaborado pela autora

Quadro 5.24 – Treinamento de Classificação com 68 instâncias para a VSM

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**47**	1	97,92%
Câncer	2	**18**	90,0%
Porcentagem Global			**95,59%**

Fonte: Elaborado pela autora

Quadro 5.25 – Matriz de confusão de teste com 29 instâncias para a VSM

	Cálculo	**Câncer**	**Porcentagem**
Cálculo	**18**	3	85,71%
Câncer	2	**6**	75,0%
Porcentagem Global			**82,76%**

Fonte: Elaborado pela autora

Conforme o resultado tem-se que a amostra com 68 instâncias de treinamento (95,59%) mostrou um desempenho um pouco melhor em relação à amostra com 83 instâncias de treinamento (83,13%).

5.2.7 Análise conjunta das técnicas

Desta forma, tem-se no quadro 5.29 a seguir, o desempenho comparativo entre as cinco técnicas analisadas para o problema do diagnóstico médico, para os dois grupos de amostras (118 e 97 instâncias).

Quadro 5.26 – Comparação do desempenho das técnicas para o caso do problema médico

Amostras	RLB		GSME-PL		FDLF		RNAs		SVM	
118	80,70%	82,86%	84,34%	80,0%	83,13%	74,29%	**100,0%**	68,57%	83,13%	80,0%
97	94,01%	79,31%	88,24%	72,41%	85,29%	75,86%	**100,0%**	72,41%	95,59%	82,76%

Fonte: Elaborado pela autora

Observe-se que os 20 resultados de treinamento e teste para as 5 técnicas tem uma diferença significativa, sendo que a diferença entre a melhor e pior acurácias é de 31,43%. De qualquer forma, para esta aplicação (Problema do diagnóstico médico), a RNAs apresenta a maior acurácia (100% tanto para treinamento como para teste).

5.3 PROBLEMA DE CALIBRAÇÃO DE BALANÇA

5.3.1 Análise Exploratória de Dados

Para o terceiro problema real abordado, problema de calibração de balança, a análise exploratória dos dados foi aplicada, conforme já mencionado, com o intuito de filtrar os dados obtendo-se, como consequência, uma maior acurácia das técnicas de DM. Para a realização dos testes foram utilizadas em duas amostras: a primeira com 1540 instâncias (644 pertencentes à classe “boa qualidade” e 896, à classe “baixa qualidade”) e a segunda com 1462 instâncias (606 pertencentes à classe “boa qualidade” e 856, à classe “baixa qualidade”). A justificativa para se trabalhar com estas duas amostras (e não apenas uma) será mostrada mais adiante.

Incialmente, verificou-se que as variáveis independentes possuem correlação através da multicolinariedade, por meio da aplicação do *software* SPSS 13.0. Os resultados encontram-se na Tabela 5.23, a seguir.

Tabela 5.23 - Coeficientes da Regressão

Variáveis	Coeficientes não padronizados		Coeficientes padronizados	t	Sig.	Estatísticas de colinearidade	
	B	Modelo padrão	Beta			Tolerância	VIF
(Constante)	0,59	0,47		1,25	0,21		
Mês	0,01	0,00	0,15	6,11	0,00	0,94	1,07
Cliente	-0,01	0,01	-0,02	-0,94	0,35	0,86	1,17
Técnico	-0,02	0,00	-0,10	-3,82	0,00	0,91	1,10
Fabricante	0,00	0,01	-0,02	-0,59	0,55	0,88	1,14
Capacidade	0,18	0,03	0,18	7,01	0,00	0,91	1,10
Corrente de Ar	0,05	0,06	0,03	0,75	0,45	0,46	2,15
Vibração	0,07	0,07	0,03	0,90	0,37	0,49	2,06
Local da Calibração	-0,01	0,03	-0,01	-0,32	0,75	0,79	1,27
Temperatura Inicial	0,00	0,00	0,04	1,49	0,14	0,75	1,34
Δ Temperatura	-0,02	0,02	-0,04	-1,14	0,25	0,63	1,59
Umidade Relativa Inicial	0,00	0,00	0,09	3,07	0,00	0,74	1,35
Δ Umidade Relativa	-0,01	0,01	-0,05	-1,65	0,10	0,63	1,59
Pressão atmosférica Inicial	0,00	0,00	-0,05	-2,06	0,04	0,93	1,07
Δ Pressão atmosférica	0,02	0,03	0,01	0,55	0,58	0,99	1,01
a. Variável dependente: Classe							

Fonte: Elaborado pela autora a partir do software SPSS 13.0

Nesta Tabela 5.23, têm-se na 1ª. coluna, a lista das variáveis; na 2ª. e 3ª. colunas, os coeficientes Beta (B) e os coeficientes padrão da Regressão; na 4ª. coluna, os coeficientes B padronizados; na 5ª. coluna, o teste *t* para cada variável; na 6ª. coluna, a Significância (Sig., ou ainda, valor *p*) e finalmente, na 7ª. e 8ª. colunas, a estatística de colinearidade, com os valores da Tolerância e do *Variance Inflation Factor* (VIF).

O SPSS não identificou nenhum atributo como tendo forte correlação com os demais (Tolerância = 0) e, portanto, não é necessário a exclusão de nenhum. Segundo Field (2009), quando o valor da Tolerância se aproxima de “0”, há forte indicação de multicolinearidade, o que diminui a qualidade do modelo. Vale ressaltar que o SPSS apresentou, automaticamente, duas tabelas, uma com as variáveis mantidas (Tabela 5.23) e outra com o diagnóstico de colinearidade, conforme a Tabela 5.24.

Tabela 5.24 – Diagnóstico de colinearidade

Variáveis	Valor próprio	Índice de condição	Proporções de variância														
			(Constante)	Mês	Cliente	Técnico	Fabricante	Capacidade	Corrente de Ar	Vibração	Local da Calibração	Temperatura Inicial	Δ Temperatura	Umidade Relativa Inicial	Δ Umidade Relativa	Pressão atmosférica Inicial	Δ Pressão atmosférica
1	13,19	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,68	4,42	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,68	0,00	0,00	0,00	0,00	0,00	0,00
3	0,32	6,44	0,00	0,01	0,65	0,00	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4	0,24	7,42	0,00	0,15	0,00	0,64	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
5	0,19	8,32	0,00	0,62	0,01	0,26	0,14	0,00	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,00	0,00
6	0,16	9,10	0,00	0,15	0,13	0,03	0,66	0,00	0,02	0,01	0,16	0,00	0,00	0,00	0,00	0,00	0,00
7	0,08	13,24	0,00	0,01	0,12	0,02	0,06	0,00	0,23	0,08	0,08	0,02	0,00	0,01	0,00	0,00	0,01
8	0,05	16,91	0,00	0,01	0,02	0,00	0,00	0,27	0,00	0,00	0,00	0,17	0,00	0,21	0,00	0,00	0,00
9	0,04	17,12	0,00	0,03	0,01	0,00	0,01	0,69	0,01	0,00	0,01	0,02	0,00	0,14	0,00	0,00	0,00
10	0,02	25,68	0,00	0,01	0,05	0,00	0,00	0,01	0,70	0,87	0,00	0,00	0,00	0,00	0,00	0,00	0,00
11	0,01	30,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,37	0,10	0,42	0,01	0,00	0,21
12	0,01	31,30	0,00	0,00	0,00	0,01	0,00	0,01	0,02	0,02	0,00	0,17	0,13	0,06	0,18	0,00	0,07
13	0,01	39,26	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,16	0,09	0,10	0,08	0,01	0,70
14	0,00	73,93	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,02	0,05	0,39	0,03	0,38	0,40	0,00
15	0,00	160,51	0,98	0,00	0,00	0,01	0,01	0,00	0,01	0,02	0,01	0,04	0,28	0,03	0,34	0,58	0,02
a. Variável dependente: Classe																	

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Já a estatística descritiva, apresentada na Tabela 5.25 a seguir, foi realizada, inicialmente, para as 1540 instâncias (644 pertencentes à classe “boa qualidade” e 896, à classe “baixa qualidade”). Esta Tabela 5.25 apresenta na 1ª. coluna a lista das variáveis; na 2ª. coluna, a quantidade de instâncias válidas (todos os 1540); na 3ª. coluna, a amplitude dos valores de cada variável; na 4ª. e 5ª. colunas, são apresentados os valores mínimo e máximo assumidos por cada variável; na 6ª. coluna, o valor médio de cada variável; na 7ª. coluna, o erro padrão da média; nas colunas 8ª. e 9ª. colunas, os desvios padrões e as variâncias, respectivamente, para cada uma das variáveis.

Tabela 5.25 - Estatística descritiva dos dados brutos das 1540 instâncias

Variáveis	N	Range	Mínimo	Máximo	Média	Modelo padrão	**Desvio padrão**	Variância
Mês	1540	24,00	1,00	25,00	13,43	0,18	**6,94**	48,18
Cliente	1540	3,00	1,00	4,00	1,64	0,03	**1,07**	1,14
Técnico	1540	7,00	1,00	8,00	4,85	0,07	**2,78**	7,73
Fabricante	1540	7,00	1,00	8,00	4,59	0,06	**2,38**	5,67
Capacidade	1540	3,00	1,00	4,00	2,05	0,01	**0,49**	0,24
Corrente de Ar	1540	1,00	0,00	1,00	0,94	0,01	**0,23**	0,05
Vibração	1540	1,00	0,00	1,00	0,91	0,01	**0,28**	0,08
Local da Calibração	1540	1,00	0,00	1,00	0,35	0,01	**0,48**	0,23
Temperatura Inicial	1540	40,20	12,80	53,00	24,34	0,11	**4,24**	17,98
Δ Temperatura	1540	18,50	0,00	18,50	8,65	0,02	**0,84**	0,71
Umidade Relativa Inicial	1540	79,90	14,10	94,00	61,72	0,30	**11,80**	139,20
Δ Umidade Relativa	1540	51,10	0,00	51,10	29,96	0,07	**2,73**	7,44
Pressão atmosférica Inicial	1540	1.024,95	0,05	1.025,00	922,65	0,97	**38,24**	1.462,66
Δ Pressão atmosférica	1540	7,50	0,00	7,50	3,47	0,01	**0,38**	0,14
Classe	1540	1,00	0,00	1,00	0,42	0,01	**0,49**	0,24
N válido (de lista)	1540							

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Pode-se observar na tabela 5.25 que os desvios padrões em parte das variáveis estão acima de “3”, ou seja, apontando a existência de dados atípicos (*outliers*), os quais deverão ser excluídos, pois poderão influenciar negativamente, piorando o desempenho das técnicas de DM. Assim, foram excluídos os dados que apresentaram os escores padronizados $z < -3$ ou $z > 3$ para cada uma das variáveis analisadas individualmente, pois esta técnica mostrou-se melhor que a distância *Mahalanobis* para o respectivo problema em questão. Foram excluídas 78 instâncias e, portanto, a amostra ficou com 1462 instâncias (606 pertencentes à classe “boa qualidade” e 856, à classe “baixa qualidade”). A Tabela 5.26 apresenta a estatística descritiva com os dados após a exclusão dos *outliers*. É possível notar, através da Tabela 5.26, que os desvios padrões das variáveis, analisados de forma conjunta, diminuíram após a exclusão dos 78 dados.

Tabela 5.26 - Estatística descritiva dos dados após a exclusão dos *outliers*

Variáveis	N	Range	Mínimo	Máximo	Média	Modelo padrão	**Desvio padrão**	Variância
Mês	1462	24,00	1,00	25,00	13,47	0,18	**6,97**	48,57
Cliente	1462	3,00	1,00	4,00	1,63	0,03	**1,07**	1,14
Técnico	1462	7,00	1,00	8,00	4,86	0,07	**2,78**	7,73
Fabricante	1462	7,00	1,00	8,00	4,60	0,06	**2,37**	5,64
Capacidade	1462	3,00	1,00	4,00	2,05	0,01	**0,50**	0,25
Corrente de Ar	1462	1,00	0,00	1,00	0,91	0,01	**0,28**	0,08
Vibração	1462	1,00	0,00	1,00	0,94	0,01	**0,24**	0,06
Local da Calibração	1462	1,00	0,00	1,00	0,36	0,01	**0,48**	0,23
Temperatura Inicial	1462	24,00	12,80	36,80	24,20	0,10	**3,90**	15,20
Δ Temperatura	1462	4,50	6,60	11,10	8,65	0,01	**0,45**	0,21
Umidade Relativa Inicial	1462	67,10	26,90	94,00	62,00	0,30	**11,48**	131,85
Δ Umidade Relativa	1462	15,20	22,10	37,30	30,07	0,04	**1,58**	2,50
Pressão atmosférica Inicial	1462	145,80	879,20	1.025,00	923,20	0,53	**20,36**	414,69
Δ Pressão atmosférica	1462	2,20	2,40	4,60	3,47	0,01	**0,28**	0,08
Classe	1462	1,00	0,00	1,00	0,41	0,01	**0,49**	0,24
N válido (de lista)	1462							

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Na sequência, o teste T^2 de Hotelling foi aplicado às duas amostras (1540 e 1462 instâncias), com a obtenção dos seguintes valores: Amostra$_{(1540\ instâncias)}$: 9,6572> 1,733 = $F_{14,1525}$ (0,95); Amostra$_{(1462\ instâncias)}$: 6,4490 > 1,733 = $F_{14,1447}$ (0,95). Por conseguinte, rejeita-se fortemente para as duas amostras, a hipótese de que as mesmas estejam centradas no mesmo vetor de médias. Assim sendo, o conjunto a calibração de balança de “boa qualidade” é distinto do de “baixa qualidade”.

Têm-se, assim, duas amostras para se aplicar a RLB: a primeira, com os dados originais, possui 1540 instâncias (644 pertencentes à classe “boa qualidade” e 896, à classe “baixa qualidade”) e a segunda, com os dados analisados estatisticamente, possui 1462 instâncias (606 pertencentes à classe “boa qualidade” e 856, à classe “baixa qualidade”).

5.3.2 Regressão Logística Binária

A técnica de DM analisada RLB, foi aplicada por meio da realização de um teste. Para o teste foi utilizada a amostra com 1462 instâncias, além disto os atributos como mês, cliente, técnico, fabricante e capacidade foram transformados em binários resultando em um total de 58 atributos. O 1º. teste se mostrou satisfatório, ou seja, suficiente para obter um desempenho aceitável da RLB, devido a quantidade de atributos.

A técnica RLB foi aplicada, inicialmente, às 1540 instâncias tanto para os dados brutos, como para os binarizados, com o auxílio do *software* SPSS 13.0

utilizando o método “Entrada Forçada” (comando “*Enter*” no SPSS), que consiste na entrada simultânea de todas as variáveis para definir o modelo final que minimiza o número de variáveis e maximiza a precisão do modelo.

A matriz de confusão de teste (Tabela 5.27) apresenta a classificação para as 1078 instâncias. A taxa de acerto global foi de 67,2% e as taxas individuais de acertos foram: para a classe “baixa qualidade”, de 81,2% e para a classe “boa qualidade”, de 47,7%. Assim, dos 627 padrões da classe “pouca qualidade”, apenas 118 estavam sendo classificados como sendo “boa qualidade” e dos 415 padrões considerados “boa qualidade”, 236 estavam sendo classificados como “pouca qualidade”, considerado insatisfatório.

Tabela 5.27 - RLB: Treinamento de Classificação para as 1078 instâncias e 14 atributos

Observado			Previsto		
			Resultado		Porcentagem
			Pouca Qualidade	Boa Qualidade	
Etapa 1	Resultado	Pouca Qualidade	**509**	118	81,2
		Boa Qualidade	236	**215**	47,7
		Porcentagem global			**67,2**

a. O valor de corte é ,500

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

No Quadro 5.30, a taxa de acerto geral é de 53,25% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 23,42% e para a classe “boa qualidade”, de 94,82%. Assim, das 269 instâncias da classe “pouca qualidade”, 206 foram erroneamente classificados como sendo da classe “boa qualidade” e das 193 instâncias consideradas como “boa qualidade”, o modelo classificou erroneamente 10 como “pouca qualidade”.

Quadro 5.27 – Matriz de confusão de teste com 462 instâncias para a RLB

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	63	206	23,42%
Boa Qualidade	10	183	94,82%
		Porcentagem Global	**53,25%**

Fonte: Elaborado pela autora

A Tabela 5.28 mostra os coeficientes B (2ª. coluna) que fazem a discriminação entre as classes. Assim, têm-se que η da equação (3.8) apresenta a forma mostrada em (5.10), a seguir, onde as variáveis X_i são as variáveis do problema (X_1 = mês; ...; X_{14} = Δ Pressão atmosférica).

Tabela 5.28 - Coeficientes da RLB considerando os 1078 instâncias e 14 atributos

	Variáveis	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Etapa 1ª	Mês	0,087	0,013	42,262	1,000	0,000	1,090	1,062	1,119
	Cliente	-0,033	0,066	0,247	1,000	0,619	0,968	0,850	1,101
	Técnico	-0,079	0,025	10,299	1,000	0,001	**0,924**	0,881	0,970
	Fabricante	0,035	0,029	1,433	1,000	0,231	1,035	0,978	1,096
	Capacidade	0,850	0,147	33,465	1,000	0,000	**2,340**	1,754	3,120
	Corrente de Ar	0,319	0,322	0,978	1,000	0,323	1,375	0,731	2,587
	Vibração	0,588	0,438	1,805	1,000	0,179	1,800	0,764	4,245
	Local da Calibração	0,184	0,159	1,335	1,000	0,248	1,201	0,880	1,640
	Temperatura Inicial	-0,012	0,017	0,540	1,000	0,462	0,988	0,955	1,021
	Δ Temperatura	-0,106	0,092	1,345	1,000	0,246	0,899	0,751	1,076
	Umidade Relativa Inicial	0,008	0,006	1,568	1,000	0,210	1,008	0,996	1,020
	Δ Umidade Relativa	-0,045	0,029	2,366	1,000	0,124	0,956	0,902	1,012
	Pressão atmosférica Inicial	-0,006	0,003	4,478	1,000	0,034	0,994	0,988	1,000
	Δ Pressão atmosférica	0,016	0,166	0,009	1,000	0,925	1,016	0,733	1,407
	Constante	4,124	3,283	1,578	1,000	0,209	61,820		

a. Variáveis inseridas na etapa 1: Mês, Cliente, Técnico, Fabricante, Capacidade, Corrente de Ar, Vibração, Local da Calibração, Temperatura Inicial, Δ Temperatura, Umidade Relativa Inicial, Δ Umidade Relativa, Pressão Atmosférica Inicial, Δ Pressão Atmosférica.

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

$$fi_{(1078)} = 0{,}087X_1 - 0{,}033X_2 - 0{,}079X_3 + 0{,}035X_4 + 0{,}850X_5 + 0{,}319X_6 + 0{,}588X_7 + 0{,}184X_8 - 0{,}012X_9 - 0{,}106X_{10} + 0{,}008X_{11} - 0{,}045X_{12} - 0{,}006X_{13} + 0{,}016X_{14} + 4{,}124 \quad (5.10)$$

Um novo teste foi aplicado na amostra com 1462 instâncias. A Tabela 5.29 de treinamento apresenta o resultado inicial considerando o modelo com apenas uma constante, ou seja, se toda a balança fosse classificado como “pouca qualidade”, a taxa de acerto seria de 58,6%, considerado insatisfatório.

Tabela 5.29 - Classificação segundo teste (1022 instâncias e 58 atributos)

Observado			Previsto		
			Resultado		Porcentagem
			Pouca Qualidade	Boa Qualidade	
Etapa 0	Resultado	Pouca Qualidade	599	0	100,0
		Boa Qualidade	424	0	0,0
	Porcentagem global				**58,6**

b. O valor de corte é ,500

Fonte: Elaborado pelos autores a partir do software SPSS 13.0

O teste Omnibus de coeficientes do modelo (ou, também chamado, de teste de ajustamento), conforme a Tabela 5.30, fornece uma indicação geral do desempenho do modelo. Observa-se que todos os valores de Sig. (Significância) estão em 0,000 (ou seja, $p < 0,0005$). Podemos concluir que o modelo com a exclusão dos 78 dados pode ser utilizado. O valor de χ^2 (chi-quadrado) é 777,511, com 53 graus de liberdade.

Tabela 5.30 - Testes de coeficientes de modelo Omnibus

		Qui-quadrado	df	Sig.
Etapa 1	Etapa	777,511	53	0,000
	Bloco	777,511	53	0,000
	Modelo	777,511	53	0,000

Fonte: Elaborado pelos autores a partir do software SPSS 13.0

Após 20 iterações na 1ª etapa, o modelo final selecionou 53 variáveis. A Tabela 5.31 mostra que na 1ª etapa o índice "R^2 de *Cox e Snell*" situou-se no patamar de 71,7%% e o "R^2 *Nagelkerke*" ficou em 53,2%. O "R^2 *Cox e Snell*" indica que 71,7% das variações ocorridas na variável dependente (baixa ou boa qualidade) são explicadas pelo conjunto das variáveis independentes, ou seja, este índice apresenta um alto índice de explicação. Da mesma forma, o índice "R^2 Nagelkerke" indica que 53,2% das variações registradas na variável dependente são explicadas pelas variáveis independentes.

Tabela 5.31 - Resumo do modelo

Etapa	Verossimilhança de log -2	R quadrado Cox & Snell	R quadrado Nagelkerke
1	610,584[a]	0,532	0,717

Fonte: Elaborado pela autora a partir do software SPSS 13.0

A Tabela 5.32 mostra que o teste "*Hosmer e Lemeshow*" também dá suporte ao modelo. Este teste, considerado por muitos como o mais confiável disponível para a avaliação do ajustamento do modelo, é interpretado de forma diferente do teste *Omnibus* anterior. Para o teste de ajustamento *Hosmer-Lemeshow*, um ajustamento pobre é indicado por um valor *p* (ou Sig) < 0,05, ou seja, para que o ajustamento seja considerado adequado o valor *p* (ou Sig) ≥ 0,05. No nosso caso o valor *p* (ou Sig) = 0,976, ou seja, o modelo proposto está suportado.

Tabela 5.32 – Teste de Hosmer e Lemeshow

Etapa	Qui-quadrado	df	Sig.
1	2,140	8	0,976

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

A Tabela 5.33 mostra os coeficientes B (2ª. coluna) para a equação (3.8) que modelará a discriminação entre as classes. Em (5.11), a seguir, têm-se a expressão de η da equação (3.8).

$$fi_{(\text{Instâncias}=1462)} = - 42{,}751X_1 – 42{,}196X_2 – 41{,}121X_3 – 42{,}478X_4 – 41{,}172X_5 – 41{,}985X_6 – 22{,}172X_7 – 22{,}004X_8 – 20{,}141X_9 – 21{,}410X_{10} - 21{,}517X_{11} – 21{,}572X_{12} – 21{,}840X_{13} – 21{,}048X_{14} – 21{,}253X_{15} - 21{,}791X_{16} – 17{,}326X_{17} + 16{,}450X_{18} + 0{,}843X_{19} + 15{,}386X_{20} +1{,}389X_{21} + 16{,}849X_{22} + 0{,}462X_{23} + 0{,}540X_{24} + 0{,}000X_{25} + 0{,}753X_{26} + 1{,}356X_{27} - 0{,}013X_{28} + 0{,}000X_{29} + 0{,}200X_{30} + 1{,}213X_{31} - 0{,}122X_{32} - 0{,}991X_{33} - 0{,}877X_{34} - 1{,}921X_{35} - 1{,}422X_{36} + + 0{,}000X_{37} + 0{,}026X_{38} – 1{,}675X_{39} + 0{,}454X_{40} + 0{,}367X_{41} – 3{,}076X_{42} + 0{,}464X_{43} – 17{,}442X_{44} + 0{,}000X_{45} – 0{,}513X_{46} + 0{,}359X_{47} + 1{,}674X_{48} + 0{,}000X_{49} + 0{,}593X_{50} - 0{,}728X_{51} + 0{,}042X_{52} + 0{,}062X_{53} - 0{,}330X_{54} + 0{,}032X_{55} - 0{,}089X_{56} - 0{,}023X_{57} - 0{,}722X_{58} + 45{,}957 \quad (5.11)$$

onde as variáveis X_i são as apresentadas na Tabela 5.33 (X_1 = mês$_1$; ...; X_{58} = Δ Pressão atmosférica), a seguir. Além disso, esta Tabela 5.33 mostra que, na verdade, as variáveis mês$_1$, mês$_2$, mês$_4$, mês$_8$, mês$_9$, mês$_{10}$, mês$_{11}$, mês$_{12}$, mês$_{13}$, mês$_{14}$, mês$_{15}$, mês$_{16}$, mês$_{17}$, mês$_{20}$, mês$_{22}$, mês$_{23}$, mês$_{24}$, técnico$_6$, fabricante$_2$, fabricante$_5$, Temperatura Inicial, Δ Temperatura, Umidade Relativa Inicial são estatisticamente significativas, fato que pode ser constatado por meio da coluna “Sig.”). Variáveis com valores inferiores a 0,05 nesta coluna são significativas.

Tabela 5.33 – Coeficientes da RLB considerando os 1462 instâncias e 58 atributos

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Etapa 1[a]	Mês(1)	-42,751	1,503E+04	0,000	1,000	0,998	0,000	0,000	
	Mês(2)	-42,196	1,537E+04	0,000	1,000	0,998	0,000	0,000	
	Mês(3)	-41,121	1,993E+04	0,000	1,000	0,998	0,000	0,000	
	Mês(4)	-42,478	1,387E+04	0,000	1,000	0,998	0,000	0,000	
	Mês(5)	-41,172	1,542E+04	0,000	1,000	0,998	0,000	0,000	
	Mês(6)	-41,985	1,453E+04	0,000	1,000	0,998	0,000	0,000	
	Mês(7)	-22,172	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(8)	-22,004	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(9)	-20,141	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(10)	-21,410	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(11)	-21,517	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(12)	-21,572	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(13)	-21,840	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(14)	-21,048	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(15)	-21,253	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(16)	-21,791	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(17)	-17,326	1,290E+04	0,000	1,000	0,999	0,000	0,000	
	Mês(18)	16,450	1,397E+04	0,000	1,000	0,999	1,393E+07	0,000	
	Mês(19)	0,843	1,380E+04	0,000	1,000	1,000	2,323	0,000	
	Mês(20)	15,386	1,433E+04	0,000	1,000	0,999	4,807E+06	0,000	
	Mês(21)	1,389	1,497E+04	0,000	1,000	1,000	4,013	0,000	
	Mês(22)	16,849	1,449E+04	0,000	1,000	0,999	2,077E+07	0,000	
	Mês(23)	0,462	1,721E+04	0,000	1,000	1,000	1,588	0,000	
	Mês(24)	0,540	1,478E+04	0,000	1,000	1,000	1,717	0,000	
	Cliente(1)	0,753	0,371	4,117	1,000	0,042	2,124	1,026	4,397
	Cliente(2)	1,356	0,709	3,661	1,000	0,056	3,883	0,968	15,580
	Cliente(3)	-0,013	0,450	0,001	1,000	0,976	0,987	0,409	2,382
	Técnico (1)	0,200	0,327	0,374	1,000	0,541	1,221	0,644	2,317
	Técnico (2)	1,213	2,456E+04	0,000	1,000	1,000	3,362	0,000	
	Técnico (3)	-0,122	0,342	0,126	1,000	0,723	0,886	0,453	1,732
	Técnico (4)	-0,991	0,860	1,326	1,000	0,250	0,371	0,069	2,005
	Técnico (5)	-0,877	0,416	4,445	1,000	0,035	0,416	0,184	0,940
	Técnico (6)	-1,921	0,752	6,521	1,000	0,011	0,147	0,034	0,640
	Técnico (7)	-1,422	0,674	4,445	1,000	0,035	0,241	0,064	0,905
	Fabricante(1)	0,026	0,459	0,003	1,000	0,955	1,026	0,417	2,523
	Fabricante(2)	-1,675	0,797	4,410	1,000	0,036	0,187	0,039	0,894
	Fabricante(3)	0,454	0,281	2,618	1,000	0,106	1,575	0,908	2,729
	Fabricante(4)	0,367	0,603	0,369	1,000	0,544	1,443	0,442	4,708
	Fabricante(5)	-3,076	1,718	3,205	1,000	0,073	0,046	0,002	1,339
	Fabricante(6)	0,464	0,892	0,270	1,000	0,603	1,590	0,277	9,131
	Fabricante(7)	-17,442	2,848E+03	0,000	1,000	0,995	0,000	0,000	
	Capacidade(1)	-0,513	1,562	0,108	1,000	0,742	0,599	0,028	12,786
	Capacidade(2)	0,359	1,172	0,094	1,000	0,760	1,431	0,144	14,248
	Capacidade(3)	1,674	1,192	1,973	1,000	0,160	**5,335**	0,516	55,168
	Corrente de Ar	0,593	0,518	1,311	1,000	0,252	1,810	0,656	4,997
	Vibração	-0,728	0,673	1,172	1,000	0,279	0,483	0,129	1,804
	Local da Calibração	0,042	0,277	0,023	1,000	0,879	1,043	0,606	1,797
	Temperatura Inicial	0,062	0,044	1,971	1,000	0,160	1,064	0,976	1,159
	Δ Temperatura	-0,330	0,278	1,407	1,000	0,236	0,719	0,417	1,240
	Umidade Relativa Inicial	0,032	0,011	8,141	1,000	0,004	1,033	1,010	1,056
	Δ Umidade Relativa	-0,089	0,080	1,232	1,000	0,267	0,915	0,783	1,070
	Pressão atmosférica Inicial	-0,023	0,006	13,167	1,000	0,000	0,977	0,965	0,989
	Δ Pressão atmosférica	-0,722	0,396	3,331	1,000	0,068	0,486	0,224	1,055
	Constante	45,957	1,290E+04	0,000	1,000	0,997	9,097E+19		

a. Variáveis inseridas na etapa 1: Mês1, Mês2, Mês3, Mês4, Mês5, Mês6, Mês7, Mês8, Mês9, Mês10, Mês11, Mês12, Mês13, Mês14, Mês15, Mês16, Mês17, Mês18, Mês19, Mês20, Mês21, Mês22, Mês23, Mês24, Cliente1, Cliente2, Cliente3, Tec1, Tec2, Tec3, Tec4, Tec5, Tec6, Tec7, Fab1, Fab2, Fab3, Fab4, Fab5, Fab6, Fab7, Cap1, Cap2, Cap3,Corrente de Ar, Vibração, Local da Calibração, Temperatura Inicial, Δ Temperatura, Umidade Relativa Inicial, Δ Umidade Relativa, Pressão Atmosférica Inicial, Δ Pressão Atmosférica.

Fonte: Elaborado pela autora a partir do *software* SPSS 13.0

Para cada valor apontado na 6ª coluna (Exp(B)) mostrado na Tabela 5.33 existe um intervalo de confiança de 95%, fornecendo um valor inferior e superior para Exp (B), também conhecido como *odds ratio*. De acordo com Tabachnick e Fidell (2013), *odds ratio* representa “a chance de estar em uma das categorias quando o valor da variável preditora (independente) aumenta em uma unidade”.

O valor *Exp (B)* é uma estimativa pontual do valor real, baseado em uma amostra. Como pode-se observar na Tabela 5.33, o preditor mais forte é o capacidade3_2, mostrando que a cada balança calibrada, aumentará em 5,335 vezes a probabilidade da balança não ser calibrada corretamente “pouca qualidade” quando o referido serviço for realizado pela capacidade$_3$. Na última coluna desta Tabela 5.33, tem-se o intervalor de confiança para estes valores: [0,516; 55,168] para a capacidade$_3$. Da mesma forma, tem-se a interpretação para os demais valores da coluna de *Exp (B)*.

As matrizes de classificação mostram uma taxa de acurácia razoavelmente satisfatória. Na Tabela 5.34, a taxa de acerto geral é de 85,5% e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 93,7% e para a classe “boa qualidade”, de 74,1,1%. Assim, das 599 instâncias da classe pouca qualidade, apenas 38 estão na classificação de boa qualidade e das 424 instâncias considerados “boa qualidade” detinham 110 instâncias “baixa qualidade”.

Tabela 5.34 - RLB: Treinamento de Classificação para as 1023 instâncias

Observado			Previsto		
			Resultado		Porcentagem
			Pouca Qualidade	Boa Qualidade	
Etapa 1	Resultado	Pouca Qualidade	**561**	38	93,7
		Boa Qualidade	110	**314**	74,1
	Porcentagem global				**85,5**

Fonte: Elaborado pelos autores a partir do *software* SPSS 13.0

No Quadro 5.31, a taxa de acerto geral é de 42,14% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 1,17% e para a classe “boa qualidade”, de 100,0%. Assim, das 257 instâncias da classe “pouca qualidade”, 254 foram erroneamente classificados como sendo da classe “boa qualidade” e das 182 instâncias consideradas como “boa qualidade”, o modelo classificou todas como “boa qualidade”.

Quadro 5.28 – Matriz de confusão de teste com 439 instâncias para a RLB

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	3	254	1,17%
Boa Qualidade	0	182	100,0%
		Porcentagem Global	**42,14%**

Fonte: Elaborado pela autora

5.3.3 Geração de uma Superfície que Minimiza Erros

Esta segunda técnica de DM constrói um modelo matemático que permite ajustar as variáveis do processo, no problema de calibração de balança, de forma a classificar a balança com “baixa qualidade” ou “boa qualidade” com um menor erro.

Foram classificadas as balanças com “baixa qualidade”, através da GSME-PL aquelas que forneceram de um valor $A_w - e_m\,\gamma + y \geq e_m$, e as balanças com “boa qualidade” aquelas que forneceram um valor $- B_w + e_k\,\gamma + z \geq e_k$.

Deste modo, utilizou-se as mesmas duas amostras (1540 e 1462 instâncias), para construir um modelo que minimiza o custo, através das variáveis do processo, com a consequente determinação do hiperplano separador das classes: $w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \quad + w_nx_n = \gamma$.

Os resultados apresentados em (5.12) e (5.13) mostram os valores da função objetivo e, também as equações para as amostras com 1078 e 1023 instâncias respectivamente. As matrizes de confusão para as duas instâncias estão apresentadas nos Quadros 5.32 e 5.35, a seguir.

Função Objetivo$_{(\text{Instâncias}=1078)}$ = 2,304

Equação minimiza erro$_{(\text{Instâncias}=1078)}$: $0{,}129X_1 - 0{,}033X_2 - 0{,}082X_3 + 0{,}055X_4 + 1{,}031X_5 + 0{,}405X_6 + 0{,}276X_7 + 0{,}134X_8 + 0{,}012X_9 - 0{,}160X_{10} + 0{,}015X_{11} - 0{,}055X_{12} - 0{,}010X_{13} - 0{,}028X_{14} = - 6{,}927$ (5.12)

A matriz de confusão para este caso está apresentada no quadro 5.13 e 5.14 a seguir.

Quadro 5.29 – Treinamento de Classificação com 1078 instâncias para a GSME-PL

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	414	213	66,03%
Boa Qualidade	177	274	60,75%
		Porcentagem Global	**63,82%**

Fonte: Elaborado pela autora

Quadro 5.30 – Matriz de confusão de teste com 462 instâncias para a GSME-PL

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**22**	247	8,18%
Boa Qualidade	5	**188**	97,41%
		Porcentagem Global	**45,45%**

Fonte: Elaborado pela autora

No Quadro 5.32, a taxa de acerto geral é de 63,82% para o treinamento e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 66,03% e para a classe “boa qualidade”, de 60,75%. No Quadro 5.33, a taxa de acerto geral é de 45,45% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 8,18% e para a classe “boa qualidade”, de 97,41%.

Função Objetivo$_{(1023)}$ = 1,455

Equação minimiza erro$_{(1023)}$: $16{,}969X_1 + 16{,}313X_2 + 16{,}313X_3 + 15{,}364X_4 + 16{,}844X_5 + 15{,}850X_6 + 17{,}988X_7 + 17{,}674X_8 + 19{,}664X_9 + 18{,}421X_{10} + 18{,}405X_{11} + 18{,}280X_{12} + 18{,}204X_{13} + 19{,}112X_{14} + 18{,}992X_{15} + 17{,}845X_{16} + 21{,}534X_{17} + 24{,}104X_{18} + 21{,}943X_{19} + 24{,}664X_{20} + 21{,}828X_{21} + 23{,}958X_{22} + 21{,}280X_{23} + 21{,}316X_{24} + 19{,}927X_{25} + 0{,}819X_{26} + 1{,}119X_{27} + 0{,}000X_{28} + 0{,}114X_{29} + 0{,}109X_{30} + 1{,}905X_{31} + 0{,}000X_{32} - 0{,}981X_{33} - 0{,}877X_{34} - 1{,}885X_{35} - 1{,}616X_{36} + 0{,}104X_{37} + 0{,}156X_{38} - 1{,}056X_{39} + 0{,}445X_{40} + 0{,}768X_{41} - 2{,}388X_{42} + 0{,}205X_{43} - 01{,}234X_{44} + 0{,}000X_{45} - 0{,}710X_{46} + 0{,}000X_{47} + 1{,}604X_{48} - 0{,}741X_{49} + 0{,}450X_{50} - 0{,}459X_{51} - 0{,}025X_{52} + 0{,}041X_{53} - 0{,}245X_{54} + 0{,}017X_{55} - 0{,}047X_{56} - 0{,}017X_{57} - 0{,}623X_{58} = 0{,}000$ (5.13)

A matriz de confusão para este caso está apresentada no quadro 5.15 e 5.16 a seguir.

Quadro 5.31 – Treinamento de Classificação com 1023 instâncias para a GSME-PL

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**534**	65	89,15%
Boa Qualidade	86	**338**	79,72%
		Porcentagem Global	**85,24%**

Fonte: Elaborado pela autora

Quadro 5.32 – Matriz de confusão de teste com 439 instâncias para a GSME-PL

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**12**	245	4,67%
Boa Qualidade	0	**182**	100,0%
		Porcentagem Global	**44,19%**

Fonte: Elaborado pela autora

No Quadro 5.34, a taxa de acerto geral é de 85,24% para o treinamento e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 89,15% e para a classe “boa qualidade”, de 79,72%. No Quadro 5.35, a taxa de acerto geral é de 44,19% para o teste e, de forma adicional, as taxas de acerto de grupos individuais foram: para a classe “pouca qualidade”, de 4,67% e para a classe “boa qualidade”, de 100,0%.

Conforme o resultado tem-se que as amostras de treinamento com 1023 instâncias (85,24%) mostraram um desempenho melhor em relação à amostra de treinamento com 1078 instâncias (63,82%).

5.3.4 Função Discriminante Linear de Fisher

A FDLF possui função discriminante $Y = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \quad + b_nX_n$, em que X_i, com i = 1; ... ; 14 para 1540 instância e 58 para 1462 instâncias representa cada uma das 14 e 58 variáveis e b_i, com i = 1; ; 14 são os coeficientes ou pesos. Desta forma, para verificar se $x_0 \in A$ ou *se* $x_0 \in B$, é necessário comparar o valor de Y com $q = ½(x_A - x_B)' S_p^{-1}(x_A - x_B)$.

Os resultados apresentados em (5.14) e (5.15) mostram as equações para as amostras com 1078 e 1023 instâncias, respectivamente. As matrizes de confusão para as duas instâncias estão apresentadas nos Quadros 5.36 e 5.39, a seguir.

$Y_{(1078)} = -4{,}781X_1 - 3{,}344X_2 - 2{,}781X_3 - 4{,}219X_4 - 2{,}844X_5 - 3{,}219X_6 - 1{,}594X_7 - 1{,}063X_8 + 1{,}281X_9 - 1{,}188X_{10} - 0{,}094X_{11} - 0{,}375X_{12} - 1{,}250X_{13} - 0{,}156X_{14} - 0{,}656X_{15} - 0{,}906X_{16} + 5{,}281X_{17} + 5{,}125X_{18} + 6{,}094X_{19} + 5{,}969X_{20} + 5{,}625X_{21} + 5{,}813X_{22} + 6{,}031X_{23} + 4{,}938X_{24} + 5{,}063X_{25} + 10{,}000X_{26} + 9{,}578X_{27} + 9{,}266X_{28} + 9{,}141X_{29} - 0{,}922X_{30} - 0{,}781X_{31} - 1{,}563X_{32} - 2{,}125X_{33} - 2{,}422X_{34} - 4{,}016X_{35} - 2{,}188X_{36} - 1{,}484X_{37} - 2{,}188X_{38} - 3{,}625X_{39} - 1{,}938X_{40} - 1{,}656X_{41} - 3{,}688X_{42} - 2{,}000X_{43} - 3{,}375X_{44} - 2{,}313X_{45} + 4{,}629X_{46} + 5{,}020X_{47} + 6{,}813X_{48} + 4{,}598X_{49} + 0{,}470X_{50} - 0{,}621X_{51} - 0{,}083X_{52} + 0{,}085X_{53} - 0{,}297X_{54} + 0{,}042X_{55} - 0{,}088X_{56} - 0{,}021X_{57} - 0{,}546X_{58} < q = -2{,}856$ e

$Y_{(1078)} = -4{,}781X_1 - 3{,}344X_2 - 2{,}781X_3 - 4{,}219X_4 - 2{,}844X_5 - 3{,}219X_6 - 1{,}594X_7 - 1{,}063X_8 + 1{,}281X_9 - 1{,}188X_{10} - 0{,}094X_{11} - 0{,}375X_{12} - 1{,}250X_{13} - 0{,}156X_{14} - 0{,}656X_{15} - 0{,}906X_{16} + 5{,}281X_{17} + 5{,}125X_{18} + 6{,}094X_{19} + 5{,}969X_{20} + 5{,}625X_{21} + 5{,}813X_{22} + 6{,}031X_{23} + 4{,}938X_{24} + 5{,}063X_{25} + 10{,}000X_{26} + 9{,}578X_{27} + 9{,}266X_{28} + 9{,}141X_{29} - 0{,}922X_{30} - 0{,}781X_{31} - 1{,}563X_{32} - 2{,}125X_{33} - 2{,}422X_{34} - 4{,}016X_{35} - 2{,}188X_{36} - 1{,}484X_{37} - 2{,}188X_{38} - 3{,}625X_{39} - 1{,}938X_{40} - 1{,}656X_{41} - 3{,}688X_{42} - 2{,}000X_{43} - 3{,}375X_{44} - 2{,}313X_{45} + 4{,}629X_{46} + 5{,}020X_{47} + 6{,}813X_{48} + 4{,}598X_{49} + 0{,}470X_{50} - 0{,}621X_{51} - 0{,}083X_{52} + 0{,}085X_{53} - 0{,}297X_{54} + 0{,}042X_{55} - 0{,}088X_{56} - 0{,}021X_{57} - 0{,}546X_{58} \geq q = -2{,}856$ (14)

As matrizes de confusão para este caso estão apresentadas nos quadros 5.36 e 5.37 a seguir.

Quadro 5.33 – Treinamento de Classificação com 1078 instâncias para a FDLF

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**400**	227	63,80%
Boa Qualidade	171	**280**	62,08%
		Porcentagem Global	**63,08%**

Fonte: Elaborado pela autora

No Quadro 5.36, têm-se que a amostra do conjunto A (pouca qualidade) pertence a A somente 63,80% e a amostra do Conjunto B (boa qualidade) pertence a B somente 62,08%.

Quadro 5.34 – Matriz de confusão de teste com 462 instâncias para a FDLF

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**29**	240	10,78%
Boa Qualidade	3	**190**	98,45%
		Porcentagem Global	**47,40%**

Fonte: Elaborado pela autora

No Quadro 5.37, têm-se que a amostra do conjunto A pertence a A somente 10,78% e a amostra do Conjunto B pertence a B somente 98,45%.

$$Y_{(68)} = -0{,}013X_1 + 1{,}151X_2 + 0{,}843X_3 - 0{,}433X_4 - 0{,}017X_5 + 0{,}007X_6 + 0{,}0008X_7 - 0{,}012X_8 + 1{,}188X_9 - 0{,}720X_{10} - 0{,}404X_{11} - 0{,}035X_{12} - 0{,}336X_{13} < q = -10{,}490 \text{ e}$$

$$Y_{(68)} = -0{,}013X_1 + 1{,}151X_2 + 0{,}843X_3 - 0{,}433X_4 - 0{,}017X_5 + 0{,}007X_6 + 0{,}0008X_7 - 0{,}012X_8 + 1{,}188X_9 - 0{,}720X_{10} - 0{,}404X_{11} - 0{,}035X_{12} - 0{,}336X_{13} \geq q = -10{,}490 \quad (5.9)$$

A matriz de confusão para este caso está apresentada no quadro 5.38 e 5.39 a seguir.

Quadro 5.35 – Treinamento de Classificação com 1023 instâncias para a FDLF

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**565**	34	94,32%
Boa Qualidade	119	**305**	71,93%
		Porcentagem Global	**85,04%**

Fonte: Elaborado pela autora

No Quadro 5.38, têm-se que a amostra do conjunto A (pouca qualidade) pertence a A somente 94,32% e a amostra do Conjunto B (boa qualidade) pertence a B somente 71,93%.

Quadro 5.36 – Matriz de confusão de teste com 439 instâncias para a FDLF

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**0**	257	0,0%
Boa Qualidade	0	**182**	100,0%
		Porcentagem Global	**41,46%**

Fonte: Elaborado pela autora

No Quadro 5.39, têm-se que a amostra do conjunto A pertence a A somente 0,0% e a amostra do Conjunto B pertence a B somente 100,0%.

Conforme o resultado tem-se que a amostra com 1023 instâncias de treinamento (85,04%) mostrou um desempenho superior em relação à amostra com 1078 instâncias de treinamento (63,08%).

5.3.5 Redes Neurais Artificiais

Esta quarta técnica de DM fez uso do algoritmo de retropropagação do erro. Assim sendo, o treinamento é realizado em duas etapas. Na primeira etapa, um padrão (m + k = 644 + 896 = 1540 instâncias) e (m + k = 606 + 856 = 1462 instâncias) pertencente aos conjuntos A ou B deste trabalho. Para o problema de calibração de balança, a RNA precisou de, aproximadamente, 100 iterações para 1078 instâncias e 100 iterações para 1462 instâncias, e assim convergir em cada uma das situações de teste (sabendo-se que o critério de convergência foi: 0,2232 para 1540 instâncias e 0,2136 para 1462 instâncias). O conjunto de dados foi dividido em três partes: 70% dos dados para treinamento e 30% para teste, usando 10 neurônios na camada oculta, pois foi o que gerou o melhor resultado entre 5, 10, 15, 20 e 25 na camada oculta.

Os resultados apresentados nas Figuras 5.6 e 5.7 apresentaram uma interpretação de treino de RNAs do *MATLAB*, onde observa-se as possibilidades de configuração de todos os parâmetros para as 1078 instâncias e 14 atributos e 1023 instâncias e 58 atributos.

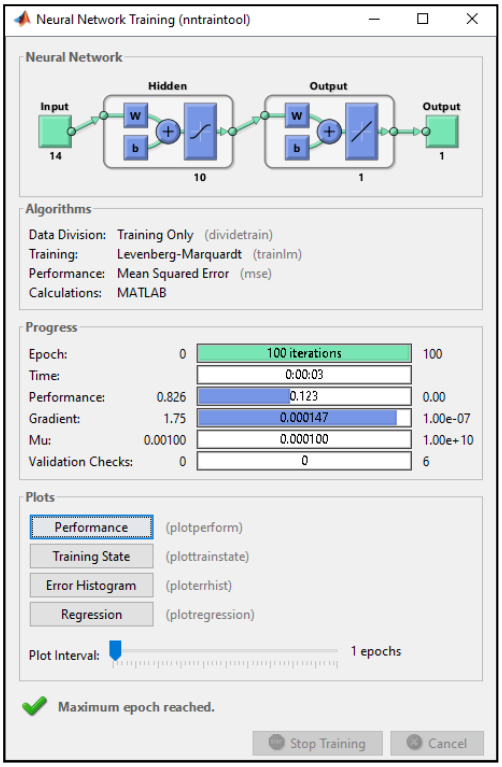

Figura 5.6 – Interpretação da RNAs com as 1078 instâncias
Fonte: Elaborado pela autora a partir do *MATLAB Starter Application*

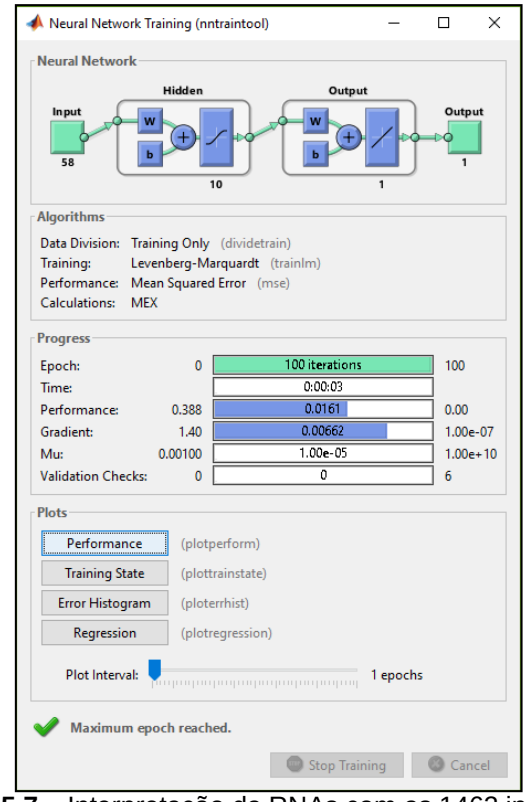

Figura 5.7 – Interpretação da RNAs com as 1462 instâncias
Fonte: Elaborado pela autora a partir do *MATLAB Starter Application*

O desempenho da rede foi medido de acordo com a média dos quadrados dos erros (MSE) e erro absoluto médio (MAE).

Teste	Treinamento
perf_mse(1078) = 0,123	perf_mse(462) = 1,024
perf_mae(1078) = 0,279	perf_mae(462) = 0,890

Teste	Treinamento
perf_mse(1062) = 0,016	perf_mse(439) = 0,721
perf_mae(1062) = 0,053	perf_mae(439) = 0,561

A matriz de confusão de treinamento e teste para estes dois casos estão apresentadas nos quadros 5.40 e 5.43 a seguir.

Quadro 5.37 – Treinamento de Classificação com 1078 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**565**	62	90,11%
Boa Qualidade	121	**330**	73,17%
		Porcentagem Global	**83,02%**

Fonte: Elaborado pela autora

Quadro 5.38 – Matriz de confusão de teste com 462 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**20**	249	7,43%
Boa Qualidade	5	**188**	97,41%
		Porcentagem Global	**45,02%**

Fonte: Elaborado pela autora

Quadro 5.39 – Treinamento de Classificação com 1023 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**597**	2	99,67%
Boa Qualidade	14	**410**	96,70%
		Porcentagem Global	**98,44%**

Fonte: Elaborado pela autora

Quadro 5.40 – Matriz de confusão de teste com 439 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**76**	181	29,57%
Boa Qualidade	3	**179**	98,35%
		Porcentagem Global	**58,09%**

Fonte: Elaborado pela autora

Conforme o resultado tem-se que a amostra com 1023 instâncias teste mostrou o mesmo desempenho em relação à amostra com 1078 instâncias treinamento (100,0%).

5.3.6 Máquina de Vetor Suporte / Support Vector Machine (SVM)

Esta quinta técnica de DM é treinada com um algoritmo baseado na teoria estatística de aprendizagem, de forma a classificar as balanças como “boa qualidade” ou “baixa qualidade” com um menor erro. Deste modo, utilizou-se a equação do hiperplano que separa o conjunto de dados (conjunto A ou B) através da equação $w^T x + b = 0$, e assim poder definir as duas regiões que contêm cada uma das classes, foi utilizado 70% para treinamento e 30% para teste.

O desempenho da rede foi medido de acordo com a média dos quadrados dos erros (MSE) e erro absoluto médio (MAE).

Teste	**Treinamento**
perf_mse(83) = 0,323	perf_mse(462) = 0,511
perf_mae(83) = 0,323	perf_mae(462) = 0,511

Teste	Treinamento
perf_mse(1062) = 0,165	perf_mse(439) = 0,585
perf_mae(1062) = 0,165	perf_mae(439) = 0,585

A matriz de confusão de treinamento e teste para estes dois casos estão apresentadas nos quadros 5.44 e 5.47 a seguir.

Quadro 5.41 – Treinamento de Classificação com 1078 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**517**	110	82,46%
Boa Qualidade	238	**213**	47,23%
		Porcentagem Global	**67,72%**

Fonte: Elaborado pela autora

Quadro 5.42 – Matriz de confusão de teste com 462 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**42**	227	15,61%
Boa Qualidade	9	**184**	95,34%
		Porcentagem Global	**48,92%**

Fonte: Elaborado pela autora

Quadro 5.43 – Treinamento de Classificação com 1023 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**559**	40	93,32%
Boa Qualidade	129	**295**	69,58%
		Porcentagem Global	**83,48%**

Fonte: Elaborado pela autora

Quadro 5.44 – Matriz de confusão de teste com 29 instâncias para a RNAs

	Pouca Qualidade	**Boa Qualidade**	**Porcentagem**
Pouca Qualidade	**0**	257	0,0%
Boa Qualidade	0	**182**	100,0%
		Porcentagem Global	**41,46%**

Fonte: Elaborado pela autora

Conforme o resultado tem-se que a amostra com 1023 instâncias de treinamento (83,48%) mostrou um desempenho superior em relação à amostra com 1078 instâncias de treinamento (67,72%).

5.3.7 Análise conjunta das técnicas

Desta forma, tem-se no quadro 5.48 a seguir, o desempenho comparativo entre as cinco técnicas analisadas para o problema da calibração da balança, para os três grupos de amostras (1540 e 1462 instâncias).

Quadro 5.45 – Comparação do desempenho das técnicas para o caso do problema da calibração da balança

Amostras	**RLB**		**GSME-PL**		**FDLF**		**RNAs**		**SVM**	
1540	67,2%	53,25%	63,82%	45,45%	63,08%	47,40%	83,02%	45,02%	67,72%	48,92%
1462	85,5%	42,14%	85,24%	44,19%	85,04%	41,46%	98,44%	58,09%	83,48%	41,46%

Fonte: Elaborado pela autora

Observe-se que os 20 resultados de treinamento e teste para as 5 técnicas tem uma diferença significativa. De qualquer forma, para esta aplicação (Problema da calibração de balança), a RNAs apresenta a maior acurácia (98,44% para treinamento com 1023 instâncias).

6 CONCLUSÕES

O presente trabalho estudou a importância da análise exploratória dos dados e algumas técnicas para DM, enquadradas no processo KDD, tendo em vista a análise dos atributos e a classificação de padrões, respectivamente. Tais técnicas foram aplicadas a três problemas reais, visando um melhor entendimento da proposta metodológica (Figura 4.1) aqui apresentada: 1) no dos cursos de PG Latu Sensu (para classificar alunos como “satisfeitos” ou “insatisfeitos”), 2) no diagnóstico médico de pacientes colestáticos (para classificar pacientes com “câncer” ou “cálculo” no duto biliar), e, também, 3) problema de calibração de balança (para classificar balanças com “boa qualidade” ou “baixa qualidade”).

Na 1ª fase do trabalho foi aplicada uma análise exploratória dos dados (teste T^2 de Hotelling; análise descritiva aos dados; descarte de dados atípicos (*outliers*); e análise dos Componentes Principais), visando a análise dos atributos e a maximização da acurácia das técnicas utilizadas na 2ª fase (RLB; GSME; FDLF; RNAs e SVM). Na 2ª fase, com a utilização das técnicas de DM, obteve-se as suas acurácias, de forma comparativa, apresentadas no Quadro 6.1, a seguir.

Quadro 6.1 – Resultado do desempenho das técnicas

Amostras		RLB		GSME-PL		FDLF		RNAs		SVM	
885	**Cursos de PG**	91,00%	**93,98%**	90,23%	90,31%	89,98%	54,51%	92,57%	90,23%	90,79%	**93,61%**
118	**Diagnóstico**	80,70%	82,86%	84,34%	80,00%	83,13%	74,29%	**100,00%**	68,57%	83,13%	80,00%
97	**Médico**	94,01%	79,31%	88,24%	72,41%	85,29%	75,86%	**100,00%**	72,41%	95,59%	82,76%
1540	**Calibração de**	67,20%	53,25%	63,82%	45,45%	63,08%	47,40%	83,02%	45,02%	67,72%	48,92%
1462	**Balança**	85,50%	42,14%	85,24%	44,19%	85,04%	41,46%	**98,44%**	58,09%	83,48%	41,46%

Para o problema dos cursos de PG Lato Sensu, com o auxílio de um especialista foram identificadas as variáveis que poderiam interferir no índice de satisfação dos alunos e, então, foi construído um questionário que foi respondido pelos alunos que participaram dos cursos de especialização a serem avaliados.

Para o problema em questão foi utilizada uma amostra com 885 instâncias. Todas as cinco técnicas apresentaram excelentes percentuais de acurácia, sendo que as técnicas que apresentaram um percentual um pouco mais elevado, foram as RLB e SVM (93,98% e 93,61%). Além disso, na Tabela 5.9, o preditor mais forte para o reporte é o Componente Principal “Docente”, com *Exp (B)* de 1.067,870, mostrando que a cada 1 aluno que avalie satisfatoriamente o componente “Docente”, diminui em 1.067 as chances dele estar insatisfeito com o curso.

Para o caso do problema médico, foram identificadas as variáveis que realmente interferem no diagnóstico médico, sendo que o médico poderá ater a tais variáveis para classificar seus pacientes. Foram utilizadas duas amostras: a 1ª com 118 instâncias e a 2ª com 97 instâncias, logo, as técnicas que apresentaram um percentual um pouco mais elevados, foram RNAs (100%).

Desta maneira, tem-se que a cada unidade a mais de “bilirrubina direta”, aumentará em 10,341 vezes a probabilidade do diagnóstico do paciente ser câncer e a cada unidade a mais de “amilase”, diminuirá em 0,219 vezes a probabilidade do diagnóstico do paciente ser câncer.

Para o problema de calibração de balança, com o auxílio de um especialista foram identificadas as variáveis que poderiam interferir no índice da qualidade dos serviços prestados, foram utilizadas duas amostras: a 1ª com 1540 instâncias e a 2ª com 1462 instâncias, logo, as técnicas que apresentaram um percentual um pouco mais elevados, foi RNAs (98,44%). Além disso, na Tabela 5.33, o preditor mais forte é o capacidade$_3$, mostrando que a cada balança calibrada, aumentará em 5,335 vezes a probabilidade da balança não ser calibrada corretamente “pouca qualidade” caso seja utilizado a capacidade$_3$.

Analisando-se estes resultados, nota-se que todas as técnicas, com a adoção da analise exploratória de dados, apresentaram uma melhora significativa nos seus desempenhos, preliminarmente à aplicação das técnicas de DM. Consequentemente é importante enfatizar a importância de se ter dados confiáveis e consistentes e, assim, dados explorados estatisticamente.

Desta forma, é possível mencionar alguns possíveis desdobramento da análise exploratória de dados. A primeira compreendeu um estudo aprofundado das técnicas e suas aplicações. A segunda compreendeu a incorporação de informações que permitissem confrontar a evolução da estrutura em relação as bases de dados estudadas. Por fim, compreendeu a realização de uma análise para avaliar a classificação de padrões.

Os métodos apresentados podem ser utilizados nos mais diversos problemas reais de classificação. Dentre as técnicas abordadas para o DM, foram utilizadas duas técnicas estatística, duas lineares e uma heurística, como sugestão seria importante utilizar novas técnicas de classificação.

Vale enfatizar que as técnicas aqui abordadas servem apenas para respaldar as decisões/conclusões do especialista, sem nunca para substitui-lo.

REFERÊNCIAS

ALES, T.V; GEVERT, G.V; CARNIERI, C; SILVA, L.C.A. Análise de crédito bancário utilizando o algoritmo sequential minimal optimization. **XLI Simpósio Brasileiro de Pesquisa Operacional – Pesquisa Operacional na Gestão do Conhecimento**, p.2242-2253, 2009.

ALZGHOUL, A.; LÖFSTRAND, M.; BACKE, B. Data stream forecasting for system fault prediction. **Computers & Industrial Engineering**, v.62, n.4, p.972-978, 2012. doi:10.1016/j.cie.2011.12.023.

ANTHERIEU, S.; AZZI, P.B.; DUMONT, J.; ABDEL-RAZZAK, Z.; GUGUEN-GUILLOUZO, C.; FROMENTY,B.; ROBIN, M.A; GUILLOUZO, A. Oxidative stress plays a major role in chlorpromazine-induced cholestasis in human heparg cells. **Hepatology**, v.57, n.4, p.1518-1529, 2013. doi: 10.1002/hep.2616.

ANUMALLA, K. **Sistema de Gestão da Pré-Processamento de Dados** (Dissertação de Mestrado). USA: Universidade de Akron, 2007.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **Requisitos gerais para a competência de laboratórios de ensaio e calibração**. Rio de Janeiro: ABNT, 2005. (ABNT ISO/IEC 17025:2005).

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **Sistemas de gestão da qualidade - Requisitos**. Rio de Janeiro: ABNT, 2008. (ISO 9001:2008).

BAPTISTELLA, A.; CUNICO, L.H.B.; STEINER, M.T.A. O uso de redes neurais na engenharia de avaliações: determinação dos valores venais de imóveis urbanos. **Revista Ciências Exatas e Naturais**, v.9, n.2, p.215-229, 2007.

BARROS, A.S.; CAMPOS, F.C. Uma discussão sobre a aplicação de processo de KDD e técnicas de mineração de dados na indústria automobilística. **XIII SIMPEP** - Bauru, SP, 2006.

BATISTA, B.C.F. **Soluções de equações diferenciais usando redes neurais de múltiplas camadas com os métodos da descida mais íngreme e Levenberg-Marquardt**. (Dissertação de Mestrado). Belém, PA: Universidade Federal do Pará, 2012.

BAZARAA, M.S.; JARVIS, J.J.; SHERALI, H.D. **Linear Programming and Networks Flows**. Fourth edition. New York: John Wiley & Sons, 2009.

BENNETT, K.P.; MANGASARIAN, O.L. Robust linear programming discrimination of two linearly inseparable sets. **Optimization Methods and Software**, v.1, p.23-34, 1992.

BERKHIN, P. **Survey of clustering data mining techniques**. Technical report, Accrue Software, San Jose, CA,2002.

BETTIOLLO, L.J.; STEINER, M.T.A. Aplicação de técnicas de reconhecimento de padrões para a investigação de Síndrome de Down no primeiro trimestre de gravidez. **Revista Ciências Exatas e Naturais**, v.11, n.2, p.265-287, 2009.

BINA, B.; SCHULTE, O.; CRAWFORD, B.; QIAN. Z.; XIONG, Y. Simple decision forests for multi-relational classification. **Decision Support Systems**, v.54, p.1269-1279, 2013. doi:10.1016/j.dss.2012.11.017.

BRAGA, A.de P.; CARVALHO, A.P.de L. F.; LUDERMIR, T.B. **Redes Neurais** artificiais: teoria e aplicações. 2. ed. Rio de Janeiro: LTC, 2011.

CABENA, P.; HADJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovering Data Mining:** from concept to implementation. New Jersey: Prentice Hall, 1998.

CAMPBELL. C. **An introduction to kernel methods**. In R. J. Howlett and L. C. Jain, editors, Radial Basis Function Networks: Design and Applications, Springer Verlag, Berlin, p.155-192, 2000.

CARDOSO, O.N.P.; MACHADO, R.T.M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de Administração Pública**, v.42, n.3, p.495-528, 2008.

CARMONA, C.J.; LUENGO, J.; GONZÁLEZ, P.; JESUS, M.J.D. An analysis on the use of pre-processing methods in evolutionary fuzzy systems for subgroup discovery. **Expert Systems with Applications**, v.39, p.11404-11412, 2012. http://dx.doi.org/10.1016/j.eswa.2012.04.029.

CARVALHO, L.A.V. **Data Mining**: mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo: Érica, 2001.

CHOUDHARY, A.K.; HARDING, J.A.; TIWARI, M.K. Data mining in manufacturing: a review based on the kind of knowledge. **Journal Intelligent Manufacturing**, v.20, n.5, p.501-521, 2009. doi: 10.1007/s10845-008-0145-x.

CIOS, K.P.; PEDRYCZ, W.; SWINIARSKI, R.W.; KURGAN, L.A. **Data Mining**: a knowledge discovery approach. Springer, New York, NY, USA, 2007.

COSTA, A.F.B.; MACHADO, M.A.G. Bivariate control charts with double sampling. **Journal of Quality Technology**, v.35, n.7, p.809-822, 2008. doi: 10.1080/02664760802061939.

COUSSEMENT, K.; VAN DEN BOSSCHE, F.A.M.; DE BOCK, K.W. Data accuracy's impact on segmentation performance: benchmarking RFM analysis, logistic regression, and decision trees. **Journal of Business Research**, v.67, p.2751-2758, 2014. http://dx.doi.org/10.1016/j.jbusres.2012.09.024.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines and other kernel-based learning methods**. Cambrigde: Cambrigde University Press, 2002.

DATTA, S.; BHADURI, K.; GIANNELLA, C.; WOLFF, R.; KARGUPTA H. Distributed Data Mining in Peer-to-Peer Networks. **IEEE Internet Computing special issue on Distributed Data Mining**, v.10, n.4, p.18-26, 2006.

DIAMANTINI, C.; POTENA, D.; STORTI, E. A virtual mart for knowledge discovery in databases. **Information Systems Frontiers**, v.15, p.447-463, 2013. doi: 10.1007/s10796-012-9399-0.

ENGEL, T.A.; CHARÃO, A.S.; KIRSCH-PINHEIRO, M.; STEFFENEL, L.A. Performance improvement of data mining in Weka through GPU acceleration. **Procedia Computer Science**, v.32, p.93-100, 2014. http://dx.doi.org/10.1016/j.procs.2014.05.402.

EROHIN, O.; KUHLANG, P.; SCHALLOW, J.; DEUSE, J. Intelligent utilization of digital databases for assembly time determination in early phases of product emergence. **Procedia CIRP**, v.3, p.424-429, 2012. http://dx.doi.org/10.1016/j.procir.2012.07.073.

FANG, X.; RACHAMADUGU, R. Policies for knowledge refreshing in databases. **Omega**, v.37, p.16-28, 2009. http://dx.doi.org/10.1016/j.omega.2006.07.003.

FAUSETT, Laurene V. **Fundamentals of neural networks**: architectures, algorithms, and applications. 1. ed. USA: Prentice Hall, 1994. 461p.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery & Data Mining**. 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 1996a.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMITH, P. The KDD process for extracting useful knowledge from volumes of data. **Comunications of the ACM**, v.39, p.27-34, 1996b.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **Artificial Intelligence Magazine**, v.17, n.3, p.37-54, 1996c.

FIELD A. **Descobrindo a Estatística usando o SPSS**. 2. Ed., Porto Alegre: Artmed, 2009.

FISHER, R.A. The use of multiple measurements in taxonomic problems**. Annals of Eugenics**, v.7, n.2, p.179-188, 1936.

FONSECA, S. O.; NAMEN, A. A. Mineração em bases de dados do INEP: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. **Educação em Revista Belo Horizonte**, v.32, n.1, p.133-157, 2016. http://dx.doi.org/10.1590/0102-4698140742.

GAGLIARDI, F. Instance-based classifiers applied to medical databases: diagnosis and knowledge extraction. **Artificial Intelligence in Medicine**, v.52, p.123-139, 2011. http://dx.doi.org/10.1016/j.artmed.2011.04.002.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining um Guia Prático.** Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. Ed. Campus, Rio de Janeiro, 2005.

GOUVÊA, M.A.; GONÇALVES, E.B.; MANTOVANI, D.M.N. Aplicação de regressão logística e algoritmos genéticos na análise de risco de crédito. **Revista Universo Contábil**, v.8, n.2, p.84-102, 2012. doi:10.4270/ruc.2012214.

GRANATYR, J. **Descoberta de regras de classificação utilizando análise formal de conceitos**. (Dissertação de Doutorado). Curitiba, PR: Pontifícia Universidade Católica do Paraná, 2011.

GUELPELI, M.V.C. **Cassiopeia:** um modelo baseado em sumarização e aprendizado autônomo usado em agrupamentos para descoberta de conhecimento em bases textuais. (Tese de Doutorado). Niterói, RJ: Universidade Federal Fluminense, 2009.

GUO, D.L.; HU, H.Y.; YI, J.Q. Neural network control for a semi-active vehicle suspension with a magnetorheological damper. **Journal of Vibration and Control**, v.10, n.3, p.461-471, 2004. doi: 10.1177/1077546304038968.

GURULER, H.; ISTANBULLU, I.; KARAHASAN, M. A new student performance analysing system using knowledge discovery in higher educational databases. **Computers & Education**, v.55, p.247-254, 2010. http://dx.doi.org/10.1016/j.compedu.2010.01.010.

HAIR, J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E.; TATHAM, R.L. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

HAN, J.; KAMBER M.; PEI, J. **Data Mining**: concepts and techniques. 3nd ed. Amsterdam; Boston: Elsevier: Morgan Kaufmann, 2011.

HARRISON, T.H. **Intranet Data Warehouse**. Berkeley, Brazil, 1998.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**: data mining, inference and prediction, 2nd ed., Springer Series in Statistics, New York, USA, 2009.

HASUMI, D.; KAMIOKA, E. A considerate application prediction system with artificial neural network. **Computer Science**, v.35, p.1547-1556, 2014. http://dx.doi.org/10.1016/j.procs.2014.08.238.

HAYKIN, S. **Neural networks and learning machines**. 3rd ed. Upper Saddle River, New Jersey: Pearson Education Inc., 2009.

HAYKIN, S. **Redes Neurais:** princípios e prática. 2. ed. Porto Alegre: Bookman, 2001.

HE, D.; GRIGORYAN, A. Multivariate multiple sampling charts. **Institute of Industrial Engineers Transactions**, v.37, p.509-521, 2005. doi:10.1080/07408170490507837.

HENNING, E.; ARAUJO, N.G; ALVES, C.C.; ZVIRTES, L. Aplicação de gráficos de controle multivariados baseados na projeção de dados via Análise de Componentes Principais. **Revista Eletrônica Produção em Foco**, v.1, p.41-66, 2011.

HINES, W.W. **Probabilidade e estatística na engenharia.** 4. Ed., tradução de Vera Regina Lima de Farias e Flores, Rio de Janeiro, 2006.

HOLM, H.; KORMAN, M.; EKSTEDT, M. A Bayesian network model for likelihood estimations of acquirement of critical software vulnerabilities and exploits. **Information and Software Technology**, v.58, p.304-318, 2015. http://dx.doi.org/10.1016/j.infsof.2014.07.001.

HOSMER, D.W.; LEMESHOW, S. **Applied logistic regression**. New York: Wiley & Sons, 2000.

HOTELLING, H. **Multivariate Quality Control**. Techniques of statistical Analysis, C. Eisenhart, M. Hastay e W. A. Wallis (eds), New York: McGraw-Hill, 1947.

INMETRO, “Vocabulário Internacional de Metrologia - Conceitos fundamentais e gerais e termos associados”. Disponível em: < http://www.inmetro.gov.br/laboratorios/termoref.asp>. Acesso em: dez./ 2016.

IOANNOU, Z.; MAKRIS, C.; PATRINOS, G.P.; TZIMAS, G. A set of novel mining tools for efficient biological knowledge discovery. **Artificial Intelligence Review**, v.42, p.461-478, 2013. doi: 10.1007/s10462-013-9413-z.

IVANCIUC, O. Applications of Support Vector Machines in Chemistry. **Reviews in Computational Chemistry**, v.23, p.291-400, 2007.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall, 2002.

JOMBART, T.; PONTIER, D.; DUFOUR, A.B. Genetic markers in the playground of multivariate analysis. **Heredity**, n.102, p.330-341, 2009. doi:10.1038/hdy.2008.130.

KAMSU-FOGUEM, B.; RIGAL, F.; MAUGET, F. Mining association rules for the quality improvement of the production process. **Expert Systems with Applications**, v.40, p.1034-1045, 2013. doi:10.1016/j.eswa.2012.08.039.

KARRAY, M.; CHEBEL-MORELLO B.; ZERHOUNI, N. PETRA: Process Evolution using a TRAce-based system on a maintenance platform. **Knowledge-Based Systems**, v.68, p.21-39, 2014. http://dx.doi.org/10.1016/j.knosys.2014.03.010.

KENDALL, M. **Multivariate Analysis**. High Wycombe. Charles Griffin, 1980.

KRISHNAMURTHY, G.T.; KRISHNAMURTHY, S. **Nuclear Hepatology**: a textbook of hepatobiliary diseases. New York, NY: Springer, 2009.

KUO, R.J.; SYU, Y.J.; CHEN. Z.Y.; TIEN, F.C. Integration of particle swarm optimization and genetic algorithm for dynamic clustering. **Information Sciences**, v.195, p.124-140, 2012. doi:10.1016/j.ins.2012.01.021.

KUSAKABE, T.; ASAKURA, Y. Behavioural data mining of transit smart card data: A data fusion approach. **Transportation Research Part C**, v.46, p.179-191, 2014. http://dx.doi.org/10.1016/j.trc.2014.05.012.

LAROSE, D.T. **Discovering Knowledge in Data:** an introduction to Data mining. John Wiley & Sons, 2003, Lawrence Erlbaum Associates, Inc., 2005.

LARROSA, A.P.Q.; MUSZINSKI, P.; PINTO, L.A.A.; Programação linear para formulação de pasta de vegetais e operação de secagem em leito de jorro. **Ciência Rural**, v.41, n.11, p.2032-2038, 2011.

LEMOS, E.P.; STEINER, M.T.A.; NIEVOLA, J.C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. **Revista de Administração (São Paulo)**, v.40, n.3, p.225-234, 2005.

LIAO, S-H.; CHU, P-H.; HSIAO, P-Y. Data mining techniques and applications – A decade review from 2000 to 2011. **Expert Systems with Applications**, v.39, p.11303-11311, 2012. http://dx.doi.org/10.1016/j.eswa.2012.02.063.

LIMA, I.; PINHEIRO, C.A.M.; SANTOS, F.A.O. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2014.

LORENA, A.C.; CARVALHO, A.C.P.L.F. Uma introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, v. 14, n.2, p. 43-67, 2007.

LUQUE-BAENA, R.M.; URDA, D.; CLAROS, M.G.; FRANCO, L.; JEREZ, J.M. Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords. **Journal of Biomedical Informatics**, v.49, p.32-44, 2014. doi:10.1016/j.jbi.2014.01.006.

LYRA, W.S.; SILVA, E.C.; ARAÚJO, M.C.U.; FRAGOSO, W.D.; VERAS, G. Classificação periódica: um exemplo didático para ensinar análise de componentes principais. **Quim. Nova**, v.33, n.7, p.1594-1597, 2010. http://dx.doi.org/10.1590/S0100-40422010000700030.

MA, L.C. A two-phase case-based distance approach for multiple-group classification problems. **Computers & Industrial Engineering**, v.63, n.1, p.89-97, 2012. http://dx.doi.org/10.1016/j.cie.2012.01.019.

MAIMON, O.; ROKACH, L. **Data Mining and Knowledge Discovery Handbook**. 2nd ed., Springer, New York, 2010.

MAINARDES, E.W.; DOMINGUES, M.J.C.O.S. Atração de Alunos para a Graduação em Administração em Joinville – SC: estudo multicaso sobre os fatores relacionados ao mercado de trabalho. **Facep Pesquisa**, v.13, n.1, p.11-47, 2010.

MARTÍNEZ-DE-PISÓN, F.J.; SANZ, A.; MARTÍNEZ-DE-PISÓN, E.; JIMÉNEZ, E.; CONTI, D. Mining association rules from time series to explain failures in a hot-dip galvanizing steel line. **Computers & Industrial Engineering**, v.63, n.1, p.22-36, 2012. http://dx.doi.org/10.1016/j.cie.2012.01.013.

MARTÍNEZ-LÓPEZ, F.J.; CASILLAS, J. Marketing Intelligent Systems for consumer behaviour modelling by a descriptive induction approach based on Genetic Fuzzy Systems. **Industrial Marketing Management**, v.38, p.714-731, 2009. http://dx.doi.org/10.1016/j.indmarman.2008.02.003.

MENDES, K.B.; FIUZA, R.M.; STEINER, M.T.A. Diagnosis of headache using artificial neural networks. **International Journal of Computer Science Network Security**, v.10, n.7, p.172-178, 2010.

MINISTÉRIO DA EDUCAÇÃO CONSELHO NACIONAL DE EDUCAÇÃO. Disponível em: < http://portal.mec.gov.br/cne/arquivos/pdf/2003/pces232_03.pdf >. Acesso em: jun./ 2016.

MINISTÉRIO DA EDUCAÇÃO. Disponível em: <http://portal.mec.gov.br/pos-graduacao>. Acesso em: fev./ 2016.

MINITAB. Disponível em: <http://support.minitab.com/pt-br/minitab/17/>. Acesso em: abril / 2016.

MORRISON, D.F. **Multivariate statistical methods**. 2ª(nd) ed., New York, Mc Graw Hill, 1976.

NGAI, E.W.T.; HU, Y.; WONG, Y.H.; CHEN. Y.; SUN, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. **Decision Support Systems**, v.50, p.559-569, 2011. http://dx.doi.org/10.1016/j.dss.2010.08.006.

NIEMINEN, P.; PÖLÖNEN, I.; SIPOLA, T. Research literature clustering using diffusion maps. **Journal of Informetrics**, v.7, p.874-886, 2013. http://dx.doi.org/10.1016/j.joi.2013.08.004.

ORRIOLS-PUIG, A.; MARTÍNEZ-LÓPEZ, F.J.; CASILLAS, J.; LEE, N. Unsupervised KDD to creatively support managers' decision making with fuzzy association rules: a distribution channel application. **Industrial Marketing Management**, v.42, n.3, p.532-543, 2013. http://dx.doi.org/10.1016/j.indmarman.2013.03.005.

PADHY, N.; MISHRA, P.; PANIGRAHI. R. The Survey of Data Mining Applications and Feature Scope. **International Journal of Computer Science, Engineering and Information Technology**, v.2, n.3, p.43-58, 2012. doi:10.5121/ijcseit.2012.2303.

PAULI-MAGNUS, C.; MEIER, P.J. Hepatobiliary Transporters and Drug-Induced Cholestasis. **Hepatology**, v.44, n.4, p.778-787, 2006. doi: 10.1002/hep.21359.

PAULI-MAGNUS, C.; MEIER, P.J.; STIEGER, B. Genetic determinants of drug-induced cholestasis and intrahepatic cholestasis of pregnancy. **Seminars in Liver Disease**, v.30, n.2, p.147-159, 2010.

PAVANELLI, A.M.; PAVANELLI, G.; STEINER, M.T.A.; COSTA, D.M.B.; GUSMÃO, B.G. Técnicas de reconhecimento de padrões aplicadas à justiça do trabalho. **Revista Eletrônica Pesquisa Operacional para o Desenvolvimento - Sobrapo**, v.3, n.2, p.90-106, 2011.

POMPONIO, L.; LE GOC, M. Reducing the gap between experts' knowledge and data: the TOM4D methodology. **Data & Knowledge Engineering**, v.94, p.1-37, 2014. http://dx.doi.org/10.1016/j.datak.2014.07.006.

QI, Z.; ALEXANDROV, V.; SHI, Y.; TIAN, Y. Parallel Regularized Multiple-Criteria Linear Programming. **Procedia Computer Science**, v.31, p.58-65, 2014. doi:10.1016/j.procs.2014.05.245.

RAUBER, T.W. “Redes Neurais Artificiais”, Disponível em: < http://www.inf.ufes.br/~thomas/pubs/eri98.pdf >. Acesso em: abril / 2016.

RELICH, M.; MUSZYNSKI, W. The use of intelligent systems for planning and scheduling of product development projects. **Procedia Computer Science**, v.35, p.1586-1595, 2014. doi:10.1016/j.procs.2014.08.242.

REZENDE, S. O. **Sistemas Inteligentes**: fundamentos e aplicações, Barueri: Editora Manole, 2003.

RIBAS, J.R.; VIEIRA, P.R. C.Da. **Análise Multivariada com o uso do SPSS**. Rio de Janeiro: Ciência Moderna, 2011.

RODRIGUES, F.F.C. **Programação da contratação de energia considerando geração distribuída no novo modelo do setor elétrico brasileiro** (Dissertação de Mestrado). Rio de Janeiro: Universidade Federal do Rio de Janeiro, COPPE, 2006.

ROEB, E.; PURUCKER, E.; GARTUNG, C.; GEIER, A.; JANSEN, B.; WINOGRAD, R.; MATERN, S. Effect of Glutathione Depletion and Hydrophilic Bile Acids on Hepatic Acute Phase Reaction in Rats with Extrahepatic Cholestasis. **Scandinavian Journal of Gastroenterology**, v.38, n.8, p.878-885, 2003. http://dx.doi.org/10.1080/00365520310003471.

ROJAS, W.A.C.; VILLEGAS, C.J.M. Graphical representation and exploratory visualization for decision trees in the KDD process. **Social and Behavioral Sciences**, v.73, p.136-144, 2013. doi:10.1016/j.sbspro.2013.02.033.

ROSA, C.R.M.; STEINER, M.T.A.; STEINER NETO, P.J. Técnicas de mineração de dados aplicadas à um problema de diagnóstico médico. **Espacios**, v.37, n.8, p.15, 2016.

RUD, O. P. **Data Mining CookBook**: modeling data for marketing risk, and customer relationship management. New York: John Wiley & Sons, Inc., 2001.

SAITTA, S., RAPHAEL, B.; SMITH, I.F.C. Data mining techniques for improving the reliability of system identification. **Advanced Engineering Informatics**, v.19, n.4, p.289-298, 2005. http://dx.doi.org/10.1016/j.aei.2005.07.005.
SANTOS, M.F.; AZEVEDO, C.S. **Data Mining** - descoberta de conhecimento em bases de dados. Lisboa, Portugal: FCA – ed. Informática – Coleção: Sistema de informação, 2005.

SASSI, R.J. **Uma Arquitetura Híbrida para Descoberta de Conhecimento em Bases de Dados**: teoria dos *rough sets* e redes neurais artificiais mapas auto-organizáveis (Tese de Doutorado). São Paulo: Escola Politécnica da Universidade de São Paulo, 2006.

SCARATTI, D.; CALVO, M.C.M. Indicador sintético para avaliar a qualidade da gestão municipal da atenção básica à saúde. **Revista Saúde Pública**, v.46, n.3, p.446-455, 2012.

SCHÖLKOPF, B.; SMOLA, A.J. **Learning with Kernels**. MIT Press, Cambridge, MA, 2002.

SHIUE, Y-R.; GUH, R-S.; TSENG, T-Y. Study on shop floor control system in semiconductor fabrication by self-organizing map-based intelligent multicontroller. **Computers & Industrial Engineering**, v.62, n.4, p.1119-1129, 2012. http://dx.doi.org/10.1016/j.cie.2012.01.004.

SILVA, E.A.Da.; COGO, F.D.; De ALMEIDA, S.L.S.; CAMPOS, K.A.; De MORAIS, A.R. Desenvolvimento de mudas de cafeeiro Coffea arabica L sob diferentes composições de substratos. **Enciclopédia Biosfera Goiânia**, v.8, n.14, p.337-346, 2012.

SILVA, E.M.; SILVA, E.M.; GONÇALVES, V.; MUROLO, A.C. **Pesquisa Operacional:** para os cursos de Administração e Engenharia. Quarta edição. São Paulo: Atlas, 2010.

SILVA, S.C.; SBRISSIA, A.F. Análise de componentes principais entre características morfogênicas e estruturais em capim-marandu sob lotação contínua. **Ciência Rural, Santa Maria***,* v.40, n.3, p.690-693, 2010.

SIM, K.; GOPALKRISHNAN, V.; ZIMEK, A.; CONG, G. A survey on enhanced subspace clustering. **Data Mining and Knowledge Discovery**, v.26, p.332-397, 2013. doi: 10.1007/s10618-012-0258-x.

SOTO, P.E.; GALLEGUILLOS, P.A.; SERÓN, M.A.; ZEPEDA, V.J.; DEMERGASSO, C.S.; PINILLA, C. Parameters influencing the microbial oxidation activity in the industrial bioleaching heap at Escondida mine, Chile. **Hydrometallurgy**, v.133, p.51-57, 2013. doi:10.1016/j.hydromet.2012.11.011.

SOUZA, A.M. **Monitoração e ajuste de realimentação em processos produtivos multivariados**. Tese (Doutorado Engenharia de Produção) - Universidade Federal Santa Catarina, 2000.

SOUZA, A.M.; POPPI, R.J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. **Quim. Nova**, v.35, n.1, p.223-229, 2012. http://dx.doi.org/10.1590/S0100-40422012000100039.

SPRUIT, M.; VROON, R.; BATENBURG, R. Towards healthcare business intelligence in long-term care an explorative case study in the Netherlands. **Computers in Human Behavior**, v.30, p.698-707, 2014. doi:10.1016/j.chb.2013.07.038.

STEINER, M.T.A.; CARNIERI, C. Pattern recognition in credit scoring analysis. **Investigacion Operativa**, v.8, n.1,2, p.1-10, 1999.

STEINER, M.T.A.; CARNIERI, C.; STANGE, P. Construção de um modelo matemático para o controle do processo de produção do papel industrial. **Pesquisa Operacional para o Desenvolvimento**, v.1, n.1, p.33-49, 2009.

STEINER, M.T.A.; CHAVES NETO, A.; BRAULIO, S.N.; ALVES, V. Métodos estatísticos multivariados aplicados à engenharia de avaliações. **Revista Gestão & Produção**, v.15, n.1, p.23-32, 2008.

STEINER, M.T.A.; NIEVOLA, J.C; SOMA, N.Y.; SHIMIZU, T.; STEINER NETO, P.J. Extração de regras de classificação a partir de redes neurais para auxílio à tomada de decisão na concessão de crédito bancário. **Pesquisa Operacional**, v.27, n.3, p.407-426, 2007. http://dx.doi.org/10.1590/S0101-74382007000300002.

STEINER, M.T.A.; SOMA, N.Y.; SHIMIZU, T.; NIEVOLA, J.C.; STEINER NETO, P.J. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Revista Gestão & Produção**, v.13, n.2, p.325-337, 2006. doi:org/10.1590/S0104-530X2006000200013.

STEINER, M.T.A. **Redes neurais artificiais**. 2014. 42 *slides*. Apresentação em *Power Point*.

TABACHNICK, B.G.; FIDELL, L.S. **Using Multivariate Statistics**. 6. Ed., California State University – Northridge. Boston: Pearson, 2013.

TAN, P.N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining** (Mineração de Dados). Rio de Janeiro: Ciência Moderna, 2009.

TRIPATHY, A.K.; ADINARAYANA, J.; SUDHARSAN, D.; VIJAYALAKSHMI, K.; MERCHANT, S.N.; DESAI, B. Data mining and wireless sensor network for groundnut pest/disease interaction and predictions - a preliminary study. **International Journal of Computer Information Systems and -Industrial Management Applications**, v.5, p.427-436, 2013. doi: 10.1109/WICT.2011.6141424.

TSAI, C.F.; YEH, H.F.; CHANG, J.F.; LIU, N.H. PHD: an efficient data clustering scheme using partition space technique for knowledge discovery in large databases. **Applied Intelligence**, v.33, n.1, p.39-53, 2010. doi: 10.1007/s10489-010-0239-y.

VAGIN, V.; FOMINA, M. Problem of knowledge discovery in noisy databases. **International Journal of Machine Learning and Cybernetics**, v.2, p.135-145, 2011. doi: 10.1007/s13042-011-0028-x.

VAN DE STEEG, E.; STRÁNECKÝ, V.; HARTMANNOVÁ, H.; NOSKOVÁ, L.; HREBÍCEK, M.; WAGENAAR, E.; ESCH, A.V.; DE WAART, D.R; ELFERINK, O.R.P.J.; KENWORTHY, K.E.; STICOVÁ, E.; AL-EDREESI, M.; KNISELY, A.S.; KMOCH, S.; JIRSA, M.; SCHINKEL, A.H. Complete OATP1B1 and OATP1B3 deficiency causes human Rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. **The Journal of Clinical Investigation**, v.122, n.2, p.519-528, 2012.

VARELLA, C.A.A. **Análise de Componentes Principais** - Análise Multivariada Aplicada as Ciências Agrárias. Universidade Federal Rural do Rio de Janeiro, Seropédica – RJ, 2008.

VIAENE, S.; AYUSO, M.; GUILLÉN, M.; GHEEL, D.V.; DEDENE, G. Strategies for detecting fraudulent claims in the automobile insurance industry. **European Journal of Operational Research**, v.176, n.1, p. 565–583, 2007. http://dx.doi.org/10.1016/j.ejor.2005.08.005.

VIEIRA, D.A.G.; **Rede Perceptron com Camadas Paralelas** (plp - Parallel Layer Perceptron), Tese de Doutorado, PPGEE/UFMG, Belo Horizonte, Dezembro de 2006.

WITTEN, I.H.; FRANK, E.; HALL, M.A. **Data Mining:** practical machine learning tools and techniques, 3rd ed. Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

XIAO, F.; FAN, C. Data mining in building automation system for improving building operational performance. **Energy and Buildings**, v.75, p.109-118, 2014. doi:10.1016/j.enbuild.2014.02.005.

YOO, I.; ALAFAIREET, P.; MARINOV, M.; PENA-HERNANDEZ, K.; GOPIDI, R.; CHANG, J.; HUA, L. Data mining in healthcare and biomedicine: a survey of the literature. **Journal of Medical Systems**, v.36, p.2431-2448, 2012. doi:10.1007/s10916-011-9710-5.

ZARAGOZÍ, B.; RABASA, A.; RODRÍGUEZ-SALA, J.J.; NAVARRO, J.T.; BELDA, A.; RAMÓN, A. Modelling farmland abandonment: A study combining GIS and data mining techniques. **Agriculture, Ecosystems and Environment**, v.155, p.124-132, 2012. doi:10.1016/j.agee.2012.03.019.

ZHANG, G.P. **Neural Networks for Data Mining**. In: Data mining and knowledge discovery handbook. 2nd ed. Springer, 2010.

ZHANG, S.C.; ZHANG, C.Q.; YANG, Q. Data preparation for data mining. **Applied Artificial Intelligence**, v.17, p.375-38, 2003.

ZHU, X.; DAVIDSON, I. **Knowledge discovery and data mining**: challenges and realities. New York: Hershey, 2007.

ZHUK, R.; IGNATOV, D.I.; KONSTANTINOVA, N. Concept learning from triadic data. **Procedia Computer Science**, v.31, p.928-938, 2014. http://dx.doi.org/10.1016/j.procs.2014.05.345.

ANEXO A – QUESTIONÁRIO APLICADO PARA A OBTENÇÃO DOS VALORES DAS VARIÁVEIS

Pós-Graduação - Instituição de Ensino Privada	
Avaliação Acadêmica do Professor e Estrutura de Apoio	
Professor	**.....**
Módulo	**.....**

Avaliador: Aluno

	Em relação ao Professor	**Muito Fraco**	**Fraco**	**Regular**	**Bom**	**Muito Bom**
1	Domínio do conteúdo	**2**	**4**	**6**	**8**	**10**
2	Didática e clareza na condução do módulo	**2**	**4**	**6**	**8**	**10**
3	Capacitade de despertar a motivação	**2**	**4**	**6**	**8**	**10**
4	Aderência do conteúdo à proposta do curso	**2**	**4**	**6**	**8**	**10**
5	Relacionamento do professor com os aluno	**2**	**4**	**6**	**8**	**10**
6	Planejamento e organização geral	**2**	**4**	**6**	**8**	**10**

	Em relação ao apoio	**Muito Fraco**	**Fraco**	**Regular**	**Bom**	**Muito Bom**
1	Sala de aula	**2**	**4**	**6**	**8**	**10**
2	Eureka & intranet	**2**	**4**	**6**	**8**	**10**
3	Estrutura, cantinas e banheiros	**2**	**4**	**6**	**8**	**10**
4	Tutor	**2**	**4**	**6**	**8**	**10**
5	Supervisão acadêmica	**2**	**4**	**6**	**8**	**10**
6	Coordenação do curso	**2**	**4**	**6**	**8**	**10**

Em relação ao Módulo	**Satisfeito**	**( )**	**Insatisfeito**	**( )**	

Críticas

Sugestões

APÊNDICE A – DADOS PARCIAIS DA BASE DE DADOS (885 X 12) OBTIDOS DO QUESTIONÁRIO -

Domínio do conteúdo	Didática e clareza na condução do módulo	Capacidade de despertar a motivação	Aderência do conteúdo à proposta do curso	Relacionamento do professor com os alunos	Planejamento e organização geral	Sala de aula	Eureka & intranet	Estrutura, cantinas e banheiros	Tutor	Supervisão acadêmica	Coordenação do curso	Classe
9,800	9,330	9,200	9,470	9,470	9,330	9,000	8,470	8,870	8,330	8,730	8,730	1
9,800	9,330	9,200	9,470	9,470	9,330	9,000	8,470	8,870	8,330	8,730	8,730	1
7,800	7,400	6,800	7,470	8,200	7,730	8,000	6,330	7,600	7,330	7,330	9,130	0
9,470	8,800	7,930	9,130	9,600	8,730	8,400	7,330	8,130	8,470	8,530	8,870	0
9,590	8,970	7,590	8,340	9,380	9,130	9,030	6,830	8,210	7,450	8,140	8,900	0

.....

.

.

.

10,000	9,946	9,838	9,459	10,000	9,784	8,811	7,556	8,108	8,944	9,027	9,135	1
9,568	9,081	9,081	9,459	9,622	9,243	8,389	7,459	7,946	8,706	8,629	8,889	0
9,923	9,385	9,462	9,462	9,923	9,231	8,308	7,615	8,462	8,500	8,500	8,417	0
10,000	9,600	9,100	9,400	9,500	9,200	8,300	7,700	8,600	8,632	8,500	8,500	1
8,480	8,000	7,960	8,320	8,960	8,240	7,560	7,160	7,680	8,612	8,776	8,816	0

APÊNDICE B – DADOS PARCIAIS DA BASE DE DADOS (118 X 13) OBTIDOS DO QUESTIONÁRIO – PROBLEMA MÉDICO

Idade	Sexo	Bilirrubina direta	Bilirrubina indireta	Fosfatases alcalinas	SGOT	SGPT	Tempo de atividade da protrombina	Albumina	Amilase	Creatinina	Leucócitos	Volume Globular	Classe
46	1	21,2	20,6	234	178	646,25	92	14	3,3	0,8	9	36,8	1
52	0	12,95	8,45	55	80	229,57	92	15	3,5	0,55	7,8	40,6	1
73	0	13,6	12,6	90	97	116,38	104	14	2,7	0,8	12,6	32,3	1
47	0	16,5	15,4	31	59	174,46	92	13	3	0,7	11,4	39	1
66	0	20,9	19,1	45	108	366,74	66	11	3,6	0,8	9,2	30,3	1

.....

.

.

.

45	0	10,2	2	10	20	158,62	76	13	2,2	0,7	8,9	42	0
50	0	5,69	2,41	16	29	158,62	189	13	2,4	0,8	8,4	38	0
39	1	5,09	2,86	20	27	158,62	219	12	3	3,3	10,3	33,6	0
66	1	2	2	20	53	285	76	18	2,8	1,2	10,7	44	0
29	1	0,5	0,5	104	57	370	189	15	3,5	0,8	8,3	44,2	0

APÊNDICE C – DADOS PARCIAIS DA BASE DE DADOS (1540 X 14) OBTIDOS DO QUESTIONÁRIO – PROBLEMA DE CALIBRAÇÃO DE BALANÇA

Mês	Cliente	Técnico	Fabricante	Capacidade	Corrente de Ar	Vibração	Local da Calibração	Temperatura Inicial	Δ Temperatura	Umidade Relativa Inicial	Δ Umidade Relativa	Pressão atmosférica Inicial	Δ Pressão atmosférica	Classe
1	1	1	3	2	1	1	0	20,7	9,3	67,0	27,1	926,0	3,5	1
1	1	1	3	2	1	1	0	26,0	8,6	60,0	30,1	927,0	3,5	1
1	1	1	3	2	1	1	0	24,2	8,6	54,0	30,1	927,0	3,5	0
1	1	1	3	2	1	1	0	24,2	8,6	54,0	30,1	927,0	3,5	1
1	1	1	3	2	1	1	0	28,6	8,6	53,0	30,1	927,0	3,5	0

.....

.
.
.

25	1	4	3	2	1	1	0	29,1	8,6	60,4	30,1	926,0	3,5	1
25	1	4	3	2	1	1	0	22,8	8,6	74,4	30,1	926,3	3,5	1
25	1	4	3	2	1	1	0	22,8	8,7	74,6	29,9	926,0	3,6	1
25	1	4	3	2	1	1	0	28,5	8,6	60,8	30,1	926,7	3,5	1
25	1	4	3	2	1	1	0	24,8	8,6	68,6	30,1	926,5	3,5	1

APÊNDICE D – VARIÁVEIS/ATRIBUTOS (CODIFICAÇÃO) - CALIBRAÇÃO DE BALANÇA

Variáveis / Atributos (tipo)	Variação	Quantidade de Entradas (Simples; "Dados Brutos")	Quantidade de Entradas (Binarizadas; "Dados Transformados")
1. Mês	1 a 25	1	25
2. Cliente	1 a 4	1	4
3. Técnico	1 a 8	1	8
4. Fabricantes	1 a 8	1	8
5. Corrente de Ar	Sim	1	1
	Não		
6. Vibração	Sim	1	1
	Não		
7. Local Calibração	Controlado	1	1
	Não controlado		
8. Capacidade	Balança analíticas I	1	4
	Balança semi-analíticas II		
	Balança industriais III		
	Balança industriais IV		
9. Temperatura Inicial	…	1	1
10. Δ Temperatura com ajuste	…	1	1
11. Umidade Relativa Inicial	…	1	1
12. Δ Umidade com ajuste	…	1	1
13. Pressão atmosférica Inicial	…	1	1
14. Δ Pressão com ajuste	…	1	1
TOTAL		**14**	**58**