ALDEMIR JUNGLOS

APLICAÇÃO DE *DATA MINING* EM BANCO DE DADOS DO SERVIÇO DE TRANSPLANTE DE MEDULA ÓSSEA

CURITIBA
2003

ALDEMIR JUNGLOS

APLICAÇÃO DE *DATA MINING* EM BANCO DE DADOS DO SERVIÇO DE TRANSPLANTE DE MEDULA ÓSSEA

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Métodos Numéricos em Engenharia, Programa de Pós-Graduação em Métodos Numéricos em Engenharia do Centro de Estudos Superiores em Engenha Civil, Setor de Tecnologia da Universidade Federal do Paraná.

Orientador: Prof. Dr. Paulo Afonso Bracarense Costa

CURITIBA

2003

À minha família pelo apoio e carinho.

AGRADECIMENTOS

Ao professor Dr. Paulo Afonso Bracarense Costa pela brilhante orientação neste trabalho.

Ao professor Dr. Anselmo Chaves Neto pelo constante ensinamento, incentivo e amizade.

Ao professor Dr. Alex Freitas, da Pós-graduação da PUC-PR, pelo acompanhamento ao longo de várias etapas deste trabalho.

Ao Dr. Ricardo Pasquini, chefe do Serviço de Transplante de Medula Óssea do HC-UFPR, pelo incentivo e sugestões.

Aos professores Dr. Julio Nievola do Mestrado em Informática aplicada da PUC-PR, Dra Lidice Cardina Lenz do TMO/HC, Dra. Maria Terezinha Steiner Arns do CESEC, Dr. Pedro José Steiner Neto do CEPPAD, Dr. Roberto Tadeu Raitz do ET/UFPR e Dr. Virgílio Balestro do Colégio Paranaense pelas sugestões e revisões deste trabalho.

À Heliz Regina Alves Neves do TMO/HC por colaborar no entendimento e tratamento dos dados.

Às manas Adelia e Angelica Junglos pelas valiosas sugestões.

À amiga Maristela, secretária do CESEC, pela animação de sempre.

Aos professores Dr. Vitor Alberto Kerber, Jussara Guimarães, Simone Peruso, Dr. Rômulo Sandrini Neto, Wilson Kachel e Ana Paula Padua Pires de Castro pelo apoio.

A PRHAE e HC-UFPR.

SUMÁRIO

LISTA DE SIMBOLOS E ABREVIATURAS

ABO – Tipo Sanguíneo e Rh de pacientes.
AER – Razão do Erro Atual ou Actual Error Rate.
ALÓGENO – TMO entre Pessoas não Aparentadas.
ANOVA – Análise da Variância ou Análisys of Variance.
APER – Razão do Erro Aparente ou Aparent Error Rate.
AUTÓLOGO – TMO utilizadas células do próprio paciente.
BUSSULFAN – Tipo de Condicionamento com Radioterapia.
C4.5 – Algoritmo de Quinlan.
CD34+ - TMO realizado com células do cordão umbilical.
CFA – Ciclofosfamida.
CORE – Sub pacote do pacote *WEKA*.
DECH – Idem GVHD.
E(Y) – Valor esperado, ou média aritmética da variável Y.
GVHDA – Grau da Doença do Enxerto Contra o Hospedeiro Aguda no Pós TMO.
GVHDC – Grau da Doença do Enxerto Contra o Hospedeiro Crônica no Pós TMO.
HC – Hospital de Clínicas.
HIDDEN –Neurônios artificiais escondidos.
IBK – Implementação do Classificador Vizinho Mais Próximo.
IBMTR – Registro Internacional de TMO ou *International Bone Marrow Transplantation Registry*
IMUNOSSUPRESSÃO – Tratamento da doença.
INV(W) – Inverso da Matriz W.
J4.8 – Algoritmo C4.5 Implementado em JAVA no pacote WEKA.
KARNOF – Situação Clínica do Paciente (*Karnofsky*, *Lansky*).
OER – Razão do Erro Ótimo ou *Optimal Error Rate*.
PAPA – Unidades de papa de hemácias recebidas pelo paciente.
PERCEPTRON – Rede que tem somente uma camada escondida
PLAQ – Quantidade de Plaquetas de Pacientes.
RN – Redes Neuronais ou Neuroniais.

RNA – Redes Neurais Artificiais.

SINGÊNICO – TMO realizado entre gêmeos.

THRESHOLD – Validação ou limites do algoritmo.

TMO – Transplante de Medula Óssea.

TPM – Probabilidade Total do Erro de Classificação ou Total Probability Misclassification.

V(Y) – Variância da variável Y.

WAIKATO – Universidade da Nova Zelândia.

WEKA – Pacote que executa algoritmos da *Data Mining*. Disponível em www.waikato.ac.nz.

LISTA DE TABELAS, GRÁFICOS E QUADROS.

RESUMO

Depois do surgimento da informática, grandes volumes de informações têm sido coletados e armazenados. Esta armazenagem por si só já trouxe grandes benefícios, antes não encontrados nos arquivos de papel. Porém, pode-se tirar muito mais proveito destes bancos de dados, propiciado pela Descoberta de Conhecimentos (KDD), como por exemplo, a Mineração de Dados ou *Data Mining* que permite investigar os dados à procura de padrões, muitas vezes não visíveis pela simples observância. A Descoberta de Conhecimento compreende os seguintes passos: ter conhecimento do domínio de aplicação; selecionar um conjunto de dados; remover os ruídos e estimar os dados faltantes; encontrar os algoritmos apropriados; classificação; interpretação dos conhecimentos descobertos e incorporação no processo.

Neste trabalho utilizou-se o banco de dados do Serviço de Transplante de Medula Óssea para, além da análise estatística para encontrar relações entre atributos, classificar os pacientes em grupos de rejeição: primeiramente em dois grupos (rejeição ou não rejeição do transplante de medula óssea -TMO) e depois em três grupos (rejeição em menos de cem dias, de cem dias a dois anos e em mais de dois anos). Além da classificação estatística (Análise Discriminante de Fisher e Regressão Logística) outras metodologias foram utilizadas, como por exemplo, Redes Neuronais, C4.5, *Decision Stump* e *Logit Boost*. Os pacotes *WEKA*, *STATGRAPHICS* e *MINITAB* foram utilizados para executar estas tarefas.

As regras aprendidas foram encaminhadas aos especialistas para possível incorporação no processo.

PALAVRAS-CHAVE: Data Mining e TMO.

ABSTRACT

After the emergence of computer science, great volumes of information have been collected and stored. The quick access to these computerized files became a great advantage over paper files. However, much more advantage can be taken from these databases provided by Knowledge Discovery (KDD), such as Data Mining that allows investigation of data in search of standards, often invisible through simple observance. Knowledge Discovery consists of the following steps: understanding the application domain; selecting a data set; removing noise and estimating missing data; finding the appropriate algorithms; classifying; interpreting discovered knowledge and incorporating it into the process.

This piece of work uses the database of the Bone Marrow Transplant Service in order to find the relation between attributes through statistical analysis and classify patients in rejection groups: first in two groups (rejection or non-rejection of bone marrow transplant) then in three groups (rejection in less than one hundred days, from one hundred days to two years and after two years). Besides statistical classification (Fisher Discriminant Analysis and Logistic Regression), other methodologies were used, such as Neuronal Networks, C4.5, Decision Stump and Logit Boost. WEKA, STATGRAPHICS and MINITAB packages were used in the execution of these tasks. The rules learned have been directed to the specialists for possible incorporation into the process.

KEY WORDS: Data Mining and TMO.

1 INTRODUÇÃO

1.1 O PROBLEMA

As duas últimas décadas acompanharam um aumento intenso na quantidade de dados que são armazenados em meio eletrônico. A quantidade de dados, o tamanho e o número de bancos de dados no mundo estão aumentando avassaladoramente. O valor destes dados armazenados está ligado à capacidade de extrair informações úteis que sirvam para dar suporte às decisões. Podem existir padrões ou tendências úteis e interessantes que, se descobertos, podem ser utilizados, por exemplo, para ajudar médicos a entender efeitos de um tratamento, ou para otimizar um processo de negócio em uma empresa, ou para ajudar no entendimento dos resultados de um experimento científico [DINIZ, 2000].

Um dos mais importantes processos de descoberta de conhecimentos é a tecnologia de Mineração de Dados (*Data Mining*).

Alguns textos da área, como por exemplo KENNEDY, utilizam os termos “mineração de dados” e “reconhecimento de padrões” com o mesmo significado, pois ambos se concentram na extração de informações ou relacionamento dos dados [ZANUSSO, 2001]. Dentro deste contexto, mineração de dados, que consiste em extrair conhecimentos implícitos e padrões ocultos em bases de dados, tem ganhado muita atenção em diversas áreas. Com o advento da *data warehousing*, que faz a armazenagem de grandes quantidades de dados em um local comum preparado para *data mining*, e do contínuo avanço no aumento do poder de processamento dos computadores, procura-se mediante tecnologias e ferramentas, extrair cada vez mais informações úteis dos dados.

A mineração de dados, por meio do uso de avançadas tecnologias, tem mostrado relações entre dados, antes escondidas, para buscar respostas em situações futuras, possibilitando que

gestores e pesquisadores tomem decisões baseadas em fatos registrados e não em suposições.

O método tradicional de transformar dados em conhecimento, que depende da análise manual de dados, está ficando impraticável em muitos domínios à medida que os volumes de dados crescem exponencialmente. A tecnologia atual de banco de dados permite o armazenamento e o acesso aos dados de forma eficiente e barata. Porém, se a origem de dados é da área empresarial, médica, científica ou governamental, o conjunto de dados na forma original tem pouco valor direto. O que é de valor é o conhecimento que pode ser extraído dos dados e sua utilização de forma adequada [DINIZ, 2000].

O pesquisador necessita hoje de respostas rápidas para perguntas; por exemplo: como descobrir padrões de dados e novos conhecimentos, ou como utilizar adequadamente e descobrir ligações entre eventos em bases de dados. Quando a escala de manipulação e exploração de dados cresce além da capacidade humana, as pessoas apelam à tecnologia de computador para a automatização.

O problema de extração de conhecimento de grandes bancos de dados envolve muitos passos, variando da manipulação e recuperação de dados à fundamentação matemática e a inferência estatística.

Quando questionado sobre o futuro próximo da *data mining*, em discussão no Portal Brasileiro [1999], Alex Alves Freitas citou que a tendência geral é que cada vez mais os sistemas gerenciadores de banco de dados oferecerem métodos de mineração de dados “embutidos” dentro de um sistema. Cita ainda que a aplicação bem sucedida de um método de *data mining* requer no mínimo dois especialistas: um para traduzir o significado dos dados e outro nos métodos de garimpagem de dados. Francisco de Assis, na mesma discussão, também cita que em breve a tecnologia de *data mining* estará disponível não só para grandes corporações, mas também para pequenas organizações [ASSIS, 1999].

Neste trabalho, esta metodologia é usada para encontrar relacionamentos lógicos entre variáveis que representam o conjunto de

dados de pacientes do Serviço de Transplante de Medula Óssea (TMO) do Hospital de Clínicas da Universidade Federal do Paraná, no intuito de criar parâmetros para entender o comportamento das variáveis dos pacientes transplantados, identificar relações e padrões entre os diversos tipos de exames coletados, e classificar os pacientes em grupos de rejeição.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Com base no conjunto de dados do Serviço de Transplante de Medula Óssea (TMO) do Hospital de Clínicas da UFPR, construir um conjunto de modelos e regras de classificação para inferir novos pacientes nas seguintes classes:

A. Rejeitar ou não rejeitar o transplante de medulo óssea;
B. Rejeitar em menos de cem dias, entre cem dias e dois anos ou em mais de dois anos.

A proposta e desafio deste trabalho são, portanto, fornecer modelos de alto desempenho e ambiente de resposta-rápida, dada a necessidade de maior ênfase na interação homem-computador, com o objetivo de dar suporte aos especialistas em transplante na tomada de decisões.

1.2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos para o presente trabalho são os seguintes.

A. Seleção dos atributos e dos dados no intuito de aplicar as técnicas do processo.

B. Transformação e redução dos dados utilizando técnicas do processo, tais como:
   - Cálculo de sumários estatísticos.
   - Cálculo de correlações entre variáveis.
   - Construção de modelos estatísticos para inferir dados faltantes do banco de dados.
   - Agrupamento.

C. Construção de regras de classificação utilizando outros algoritmos de classificação.

D. Avaliação e aplicação das regras aprendidas.

1.3 JUSTIFICATIVA

O Banco de Dados do Hospital de Clínicas da Universidade Federal do Paraná tem proporções volumosas, pela quantidade de pacientes e de variáveis coletadas junto aos pacientes. Têm-se aí, então, infinidades de cruzamentos e análises que se podem fazer para descobrir, por algum processo, padrões nos dados.

Com a devida organização dos dados e reconhecimento de padrões é possível a capacitação dos médicos na análise das informações nos mais diferentes registros (pacientes) e, por meio da identificação de relações, é possível planejar estratégias, aprender com erros ocorridos e aplicar corretas soluções em outras estratégias.

O médico não é especialista em descobrir informações em banco de dados, mas uma pessoa responsável por perceber o sentido das informações, pois participa de todo processo e tem como referência a bibliografia que mostra os melhores caminhos para o sucesso do transplante. Assim, desenvolver um processo usando técnicas da descoberta de conhecimento, poderia auxiliá-lo na tomada de decisões, usando as informações obtidas de grandes bancos de dados que são afins com as suas atividades.

1.4 ESTRUTURA DO TRABALHO

Além da introdução no primeiro capítulo, esta dissertação é composta de uma Revisão de Literatura no segundo capítulo, onde aborda a Classificação entre Populações. No terceiro capítulo, Material e Métodos, tem-se a população, amostra e as variáveis analisadas. No quarto capítulo apresentou-se os resultados e as discussões, e finalmente no quinto e último capítulo faz-se a conclusão.

2 REVISÃO BIBLIOGRÁFICA – A TECNOLOGIA DE MINERAÇÃO DE DADOS

2.1. INTRODUÇÃO

Talvez a definição mais importante da *Data Mining* tenha sido elaborada por Usama Fayyad (FAYYAD et al. 1996): *Data Mining* é o processo não trivial de identificar padrões válidos, potencialmente úteis e compreensíveis, em um conjunto de dados.

Minerar dados pode ser definido, então, como o processo de extrair informações válidas, previamente desconhecidas, a partir de grandes bases de dados, usando-as para efetuar decisões cruciais (Tutorial IBM, 1996).

Na busca de regras que não são perceptíveis ou não são óbvias nos bancos de dados, a escolha apropriada da ferramenta é extremamente importante, pois auxiliará na extração de novas características.

A tecnologia de mineração de dados pode identificar padrões de comportamento nos dados. O diferencial está no fato de que as descobertas de padrões se dão por uma lógica de algoritmos, que são ferramentas de descobertas matemáticas feitas sobre os registros já processados. Este processo auxilia na descoberta de informações relevantes como padrões, associações, mudanças, anomalias e estruturas em grande quantidade de dados armazenados.

Algoritmos de reconhecimento de padrões tendem a cair num problema de otimização relativamente simples como, por exemplo, o gradiente descendente, embora técnicas mais sofisticadas de otimização também tenham sido utilizadas.

Mineração de dados está baseada em várias áreas, principalmente a Estatística e a Informática.

Quanto aos modelos, há dois fatores relevantes: as características do modelo (classificação, agrupamento e associação) e a forma de representação dos modelos (regras ou árvores de decisão).

Modelos mais complexos podem-se ajustar melhor aos dados, mas também podem ser os mais difíceis de se entender. Enquanto pesquisadores tendem a defender modelos complexos, especialistas envolvidos em aplicações freqüentemente utilizam modelos mais simples principalmente pelo fato da sua generalidade e interpretabilidade. Normalmente há um critério quantitativo explícito embutido no algoritmo de busca; por exemplo, o critério de máxima probabilidade de encontrar os parâmetros. O problema de encontrar os melhores parâmetros é freqüentemente reduzido a um problema de otimização.

2.2. DESCOBERTA DE CONHECIMENTO

O termo KDD, que significa Descoberta de Conhecimento em Banco de Dados ou *Knowledge Discovery in Data* pode ser visto como o processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados [FELDENS, 2000]. Este processo se vale de tecnologias de reconhecimento utilizando padrões e técnicas estatísticas. Garimpagem em banco de dados é uma das técnicas utilizada para a realização do KDD para extrair padrões dos dados [NORTON, 1999]. O KDD evoluiu e continua evoluindo da interseção de pesquisas em campos tais como: bancos de dados, aprendizado de máquinas, reconhecimento de padrões, estatística, inteligência artificial, aquisição de conhecimento, visualização de dados, descoberta científica, recuperação de informação e computação de alto-desempenho [ZANUSSO, 2001].

O conhecimento descoberto deve ser não apenas válido, com alto grau de confiabilidade, mas também compreensível e potencialmente

útil para o usuário. Em muitos casos, é possível definir medidas de confiabilidade ou utilidade [FAYYAD et al., 1996].

O processo de descoberta de conhecimento em banco de dados é interativo e repetitivo, envolvendo vários passos, conforme descritos a seguir:

A. CONHECIMENTO DO DOMÍNIO DA APLICAÇÃO. Ter conhecimento relevante anterior e as metas da aplicação.
B. CRIAÇÃO DE UM BANCO DE DADOS ALVO. Selecionar um conjunto de dados ou dar ênfase para um subconjunto de variáveis ou exemplo de dados nos quais a "descoberta" será realizada.
C. PRÉ-PROCESSAMENTO DOS DADOS. Fazer operações básicas, como remover ruídos ou subcamadas se necessário; coletando informações necessárias para modelar; decidindo estratégias para manusear dados faltantes e decidindo assuntos como tipos de dados, esquema e mapeamento de valores desconhecidos. Este item provê um resumo conciso e sucinto de uma coleção de dados, com o objetivo de eliminar ruídos ou distorções.
D. REDUÇÃO DE DADOS. Encontrar formas práticas para representar dados. Compreende uma descrição compacta para um subconjunto de dados. Funções mais sofisticadas envolvem regras de resumo, técnicas de visualização multivariada e funções de relação entre variáveis. Funções de sumarização são também freqüentemente utilizadas em métodos interativos de exploração de análise de dados e geração de relatórios.
E. ESCOLHA DA FUNÇÃO. Encontrar modelos derivados dos algoritmos; como por exemplo, sumarização, regressão e classificação.
F. ENCONTRAR O ALGORITMO. Selecionar métodos de procura por modelos e decidir quais modelos e parâmetros podem ser apropriados.
G. MINERAÇÃO DE DADOS. Procurar por modelos de interesse numa forma particular de representação ou num conjunto de tais representações, incluindo regras de classificação ou árvores,

regressão, agrupamento, dependência e análise linear. A classificação analisa um conjunto de dados de treinamento e constrói modelos para cada classe, com base nas características dos dados. Por exemplo, uma árvore de decisão ou um conjunto de regras de classificação é gerado por tal processo, que pode ser usado para entender melhor cada classe no banco de dados e para classificar dados futuros. Por exemplo, alguém pode classificar doenças e ajudar a prever tipos de doenças, com base nos sintomas dos pacientes. Existem muitos métodos de classificação desenvolvidos no campo de aprendizagem de máquina, estatística, redes neuronais e outros. As técnicas de classificação criam automaticamente um modelo a partir de um conjunto inicial de registros. Esse conjunto é chamado de conjunto de treinamento. Os registros do conjunto de treinamento devem pertencer a um pequeno grupo de classes pré-definidas pelos analistas.

H. INTERPRETAÇÃO. Visa à explicitação do modelo descoberto, removendo modelos redundantes ou irrelevantes e traduzindo os úteis em termos compreendidos pelos usuários.

I. UTILIZAÇÃO DAS REGRAS DESCOBERTAS. Incorporar este conhecimento no processo.

2.3. A MINERAÇÃO DE DADOS E A ANÁLISE ESTATÍSTICA

Mineração de Dados pode ser vista como descendente direta da estatística (LOYOLLA, 2000).

Conceitos como distribuição normal, variância, análise de regressão, análise de dispersão dos dados, análise discriminante, análise de agrupamento, intervalos de confiança e teste de hipótese são utilizados para realizar as pesquisas nos dados, bem como analisar e descobrir relacionamentos entre eles.

Os procedimentos estatísticos de análise são muito úteis em pesquisas em banco de dados pequenos e “limpos”, coletados para responder a certa quantidade de questões particulares.

2.4. INTELIGÊNCIA ARTIFICIAL

Em 1956 no *Darthmouth College* nasceram os dois paradigmas da Inteligência Artificial: Simbólica e Conexista.

Segundo Allard e Fuchs, Inteligência Artificial pode ser definida como um conjunto de modelos, algoritmos, técnicas, ferramentas e aplicações em um sistema computadorizado, que emula algumas das habilidades cognitivas do homem.

2.5. APRENDIZAGEM DE MÁQUINA

A função do aprendizado de máquina é fazer com que, a partir do banco de dados, regras sejam “aprendidas” pelos programas, tornando-os capazes de tomarem decisões por meio de modelos que se fazem mais "inteligentes", cada vez que se acrescenta mais informação ao seu processamento.

2.6. RECONHECIMENTO DE PADRÕES

Reconhecimento de Padrões pode ser definido como um processo de identificar estruturas nos dados por comparações com estruturas conhecidas; as estruturas conhecidas são desenvolvidas através de métodos de classificação em [ROSS, 1995].

Outra definição [KLIR, 1995]: "O reconhecimento de padrões pode ser definido como um processo pelo qual buscam-se estruturas nos dados e classificam-se estas estruturas dentro de categorias tais que o grau de associação é maior entre as estruturas da mesma categoria e menor entre as categorias de estruturas diferentes. As categorias relevantes são usualmente caracterizadas por estruturas prototípicas

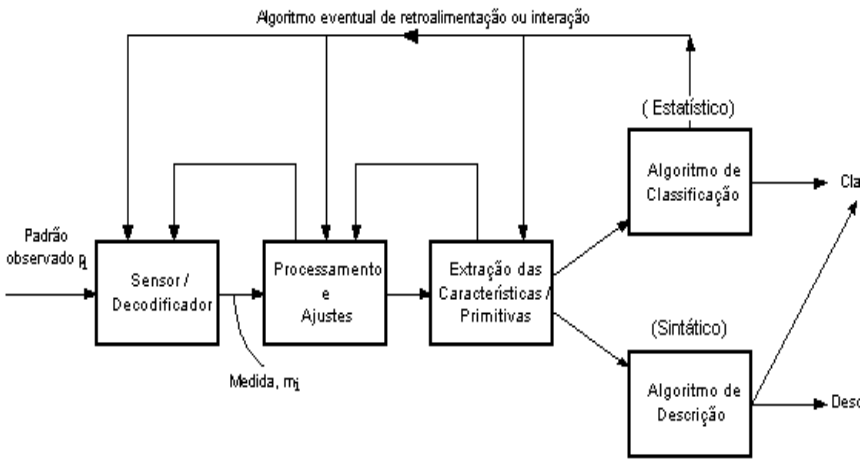
derivadas da experiência do passado. Cada categoria pode ser caracterizada por mais de uma estrutura prototípica".

A estrutura de um sistema típico de reconhecimento de padrões é mostrada no gráfico 3 na página a seguir (RAITZZ, 1997).

Segundo Bishop (1995), a forma mais geral e natural de formular soluções para o reconhecimento de padrões é o Reconhecimento Estatístico, através do qual é reconhecida a natureza estatística tanto da informação que se quer representar quanto dos resultados que devem ser expressos. Robert Schalkoff (1992) diz que o reconhecimento de padrões estatístico assume uma base estatística para os algoritmos de classificação. Um conjunto de medidas é extraído dos dados de entrada e usado para associar cada vetor de características a uma determinada classe.

Em geral, as tarefas desta tecnologia podem ser classificadas em duas categorias: descritiva e prognóstica. A primeira descreve o conjunto de dados de uma maneira concisa e resumida e apresenta propriedades gerais interessantes dos dados; a segunda constrói um conjunto de modelos, realiza inferências sobre o conjunto de dados disponíveis e tenta predizer o comportamento de novos conjuntos de dados.

GRÁFICO 3 – DIAGRAMA DA ESTRUTURA DO RECONHECIMENTO DE PADRÕES.



2.7. CLASSIFICAÇÃO ENTRE DUAS OU MAIS POPULAÇÕES

2.7.1. INTRODUÇÃO

Os algoritmos de classificação são desenhados para “aprender” e utilizar os resultados mais apropriados para a tomada de decisão. Dois tipos de regras de classificação foram utilizados neste estudo: para duas e para três populações. No caso de classificação estatística aplicou-se a Função Discriminante Linear de Fisher (para duas e três populações) e Regressão Logística (para duas populações).

Além dos métodos estatísticos de classificação, outros métodos foram utilizados neste trabalho; como por exemplo, os algoritmos:

- Redes Neuronais.
- J4.8 (C4.5 revisão 8 implementado em Java).
- Decision Stump-DS.
- IBK.
- Logit Boost.
- Zero R (probabilidades a priori).
- Naive Bayes.

Neste trabalho não houve a preocupação de tratar a fundo a teoria desses métodos.

A maioria dos algoritmos de aprendizagem de máquina é desenhada para, utilizando os atributos mais apropriados, identificar e classificar os padrões, melhorando a taxa de acerto e facilitando a interpretação dos resultados para a tomada de decisão. Continuamente, os atributos irrelevantes não são selecionados na construção das regras.

Depois de determinado o conjunto de dados, é colocado em ordem aleatória para prevenção de tendências associadas à ordem de apresentação dos dados. Essencialmente, pode ser necessário pré-processar estes dados, por meio de normalizações e conversões de

formato para torná-los mais apropriados à sua utilização. Esta tarefa requer uma análise cuidadosa do problema para minimizar ambigüidades e erros nos dados.

Quanto ao modo de treinamento, na prática é mais utilizado o modo padrão, devido ao menor armazenamento de dados, além de ser menos suscetível ao problema de mínimos locais, devido à pesquisa de natureza estocástica que realiza. Por outro lado, no modo *batch* se tem uma melhor estimativa do vetor gradiente, o que torna o treinamento estável. A eficiência relativa dos dois modos de treinamento depende do problema que está sendo tratado (WITTEN, I. H, 1999).

a) Modo Padrão ou *on-line* - a correção dos pesos acontece a cada apresentação à rede de um exemplo do conjunto de treinamento. Cada correção de pesos baseia-se somente no erro do exemplo apresentado naquela iteração. Assim, em cada ciclo ocorrem N pares (entrada e saída) de correções, onde N é o número total de exemplos.

b) Modo Agrupamento ou *Batch* - apenas uma correção é feita por ciclo. Todos os exemplos do conjunto de treinamento são apresentados à rede, seu erro médio é calculado; a partir deste erro, fazem-se as correções dos pesos.

Quanto ao tempo de treinamento, vários fatores podem influenciar a sua duração; porém sempre será necessário utilizar algum critério de parada. Podem ser consideradas a taxa de erro médio por ciclo e a capacidade de generalização.

O conjunto de testes é utilizado para determinar a performance do algoritmo com dados que não foram previamente utilizados. A performance, medida nesta fase, é boa indicação de seu desempenho real. (WITTEN & FRANK, 1999).

A seguir são descritos alguns algoritmos de aprendizagem utilizados neste trabalho.

2.7.2. METODOLOGIAS ESTATÍSTICAS DE CLASSIFICAÇÃO ENTRE DUAS OU MAIS POPULAÇÕES

A. METODOLOGIA DE FISHER PARA CLASSIFICAÇÃO ENTRE DUAS POPULAÇÕES

A idéia de Fisher foi transformar as observações multivariadas $\underset{\sim}{X}$'s em observações univariadas *Y's*, de forma que as populações Π_1 e Π_2 sejam separadas tanto quanto possível.

Sejam μ_{iy}, $i = 1, 2,$ as médias dos escores Y´s obtidos dos vetores originais $\underset{\sim}{X}$'s pertencentes às populações Π_i, $i = 1, 2$. Então, Fisher selecionou a combinação linear que maximiza a distância quadrática entre as médias μ_{iy}, relativamente à variabilidade dos Y´s (JOHNSON R.A, 1998). Seja $\underset{\sim}{X}_i$ os valores esperados da observação multivariada de cada população, conforme [3.1] a seguir:

$$\underset{\sim}{\mu}_i = E[\underset{\sim}{X}_i \mid \Pi_i], \quad i = 1, 2. \qquad [3.1]$$

Supondo que a matriz de covariância Σ seja a mesma para ambas as populações, ou seja,

$$\Sigma = E[\underset{\sim}{X} - \underset{\sim}{\mu}_i][\underset{\sim}{X} - \underset{\sim}{\mu}_i]' \qquad i = 1, 2, \qquad [3.2]$$

e considerando a combinação linear $Y = \underset{\sim}{c}'\underset{\sim}{X}$, tem-se a seguinte expressão para a média do escore Y,

$$\underset{\sim}{\mu}_{iY} = E[Y \mid \Pi_i] = E[\underset{\sim}{c}'\underset{\sim}{X} \mid \Pi_i] = \underset{\sim}{c}'\underset{\sim}{\mu}_i, \quad i = 1,2. \qquad [3.3]$$

Analogamente, a variância de Y, conforme [3.4], é dada por:

$$V(Y) = \sigma_Y{}^2 = V(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}'V(\underset{\sim}{X})\underset{\sim}{c} = \underset{\sim}{c}'\Sigma\underset{\sim}{c} \qquad [3.4]$$

Segundo Fisher, a melhor combinação linear vem da razão entre o quadrado da distância entre as médias e a variância de Y, ou seja,

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{(\underset{\sim}{c}'\delta)^2}{\underset{\sim}{c}'\Sigma\underset{\sim}{c}}, \qquad \text{onde} \underset{\sim}{\delta} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \qquad [3.5]$$

que é maximizada quando se tem $\underset{\sim}{c} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\Sigma^{-1}\underset{\sim}{X}$. Então a Função Discriminante Linear de Fisher que é dado por:

$$Y = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{X} \qquad [3.6]$$

Esta função transforma populações multivariadas Π_i, i=1, 2 em populações univariadas, de modo que as médias das populações univariadas são separadas tanto quanto possível. Assim, tomando-se y_o como o valor da função discriminante de Fisher para uma nova observação $\underline{x}_o$, tem-se conforme [3.7] a seguir o Valor do escore univariado correspondente ao vetor $\underline{x}_o$,

$$y_o = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}_0 \qquad [3.7]$$

Agora, considerando o ponto médio *m* entre as médias das duas populações univariadas,

$$m = \frac{\mu_{1Y} + \mu_{2Y}}{2} = \frac{(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)}{2}, \qquad [3.8]$$

tem-se por [3.9] a regra de classificação:

$$\begin{cases} E(y_0 \mid \pi_1) - m \geq 0 \;\; \Rightarrow \;\; \underset{\sim}{x}_0 \in \pi_1 \\ E(y_0 \mid \pi_2) - m < 0 \;\; \Rightarrow \;\; \underset{\sim}{x}_0 \in \pi_2 \end{cases} . \qquad [3.9]$$

Assim, se o vetor $\underline{x}_o$ da nova observação pertence à população π_1, espera-se que o escore y_o seja igual ou maior do que o ponto médio *m*; se $\underline{x}_o$ pertence à população π_2, espera-se que o escore y_o seja menor do que o ponto *m*. Portanto, a regra de classificação obedece, conforme [3.10], ao seguinte critério:

$$\text{Alocar } \underset{\sim}{x}_o \text{ em } \begin{cases} \pi_1, \text{ se } (y_o - m) \geq 0 \\ \pi_2, \text{ se } (y_o - m) < 0 \end{cases} \qquad [3.10]$$

Porém, como os parâmetros populacionais (médias e covariâncias) são geralmente desconhecidos, utiliza-se o resultado amostral.

Sejam os vetores amostrais obtidos em amostras de tamanhos n_1 para a população Π_1 e n_2 para a população Π_2 observações. Então, as médias amostrais são dadas por:

$$\begin{cases} \underset{\sim 1}{\bar{x}} = \dfrac{1}{n_1} \displaystyle\sum_{i=1}^{n_1} \underset{\sim}{x}_{i1} \\ \underset{\sim 2}{\bar{x}} = \dfrac{1}{n_2} \displaystyle\sum_{i=1}^{n_2} \underset{\sim}{x}_{i2} \end{cases} \qquad [3.11]$$

e as covariâncias amostrais,

$$\begin{cases} S_1 = \dfrac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\underset{\sim}{x}_{i1} - \underset{\sim}{\bar{x}}_1)(\underset{\sim}{x}_{i1} - \underset{\sim}{\bar{x}}_1)' \\ S_2 = \dfrac{1}{n_2 - 1}\sum_{i=1}^{n_2}(\underset{\sim}{x}_{i2} - \underset{\sim}{\bar{x}}_2)(\underset{\sim}{x}_{i2} - \underset{\sim}{\bar{x}}_2)' \end{cases} \qquad [3.12]$$

Logo, a matriz de covariância comum Σ é estimada pela conjunta das amostras S_p, conforme a seguir:

$$S_p = \hat{\Sigma} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}, \qquad [3.13]$$

que é um estimador não viciado.

Assim, a Função Discriminante Linear de Fisher Amostral é dada em [3.14] por:

$$y = \underset{\sim}{\hat{c}}'\underset{\sim}{x} \qquad \text{onde} \quad \underset{\sim}{\hat{c}} = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)'\, S_p^{-1} \qquad [3.14]$$

e a estimativa do ponto médio *m* entre a média das duas populações univariadas é dada por:

$$\hat{m} = \frac{1}{2}\left[\bar{y}_1 + \bar{y}_2\right] = \frac{1}{2}\left[(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)'S_p^{-1}(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2)\right] \qquad [3.15]$$

$$\bar{y}_1 = \underset{\sim}{\hat{c}}'\underset{\sim}{\bar{x}}_1 \quad \text{e} \quad \bar{y}_2 = \underset{\sim}{\hat{c}}'\underset{\sim}{\bar{x}}_2 \qquad [3.16]$$

Portanto, a regra de classificação obedece ao critério conforme colocado em [3.17] a seguir:

$$\text{Alocar } \underset{\sim}{x}_o \text{ em } \begin{cases} \pi_1, \text{ se } y_0 = [(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \, S_p^{-1} \, \underset{\sim}{x}_0 - \hat{m}] \geq 0 \\ \pi_2, \text{ se } y_0 = [(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \, S_p^{-1} \, \underset{\sim}{x}_0 - \hat{m}] < 0 \end{cases} \qquad [3.17]$$

B. METODOLOGIA DE FISHER PARA CLASSIFICAÇÃO ENTRE DIVERSAS POPULAÇÕES

Fisher propôs a extensão do seu método de classificação exposto anteriormente para diversas populações. Neste caso, tem-se como objetivo encontrar regras de classificação para alocar novos indivíduos em uma das *g* populações Π_i, $i = 1,2,3,...,g$ [JOHNSON, 1998].

- A motivação da análise discriminante de Fisher é a necessidade de se obter representações de populações que envolvam poucas combinações lineares das observações [CHAVES NETO, 2001].

Este método não exige a suposição de que as populações sejam normais multivariadas. Entretanto assume-se que as matrizes de covariâncias Σ sejam iguais e com posto completo, ou seja, $\Sigma = \Sigma_1 = \Sigma_2 = \Sigma_3 = ... = \Sigma_g$, onde g é o número de populações.

Seja $\bar{\underset{\sim}{\mu}}$ o vetor médio de diversas populações (grupos) calculado em [3.18] da seguinte maneira:

$$\bar{\underset{\sim}{\mu}} = \frac{\sum_{i=1}^{g} \underset{\sim}{\mu}_i}{g}. \qquad [3.18]$$

Seja B_o a matriz das somas dos produtos cruzados entre grupos populacionais, conforme se esclarece em [3.19].

$$B_o = \sum_{i=1}^{g} (\underset{\sim}{\mu}_i - \underset{\sim}{\bar{\mu}})(\underset{\sim}{\mu}_i - \underset{\sim}{\bar{\mu}})' \qquad [3.19]$$

onde $\bar{\mu} = \dfrac{\sum_{i=1}^{g} \mu_i}{g}$

Seja Y a combinação linear com esperança E(Y) e variância V(Y) para a população Π_i, $i = 1, 2, 3, ..., g$, dada em [3.20]. Seja a combinação linear Y,

$$Y = \underset{\sim}{c}' \underset{\sim}{x}, \qquad [3.20]$$

com seu valor esperado E(Y),

$E(Y) = \underset{\sim}{c}' E(\underset{\sim}{x} \mid \pi_i) = \underset{\sim}{c}' \underset{\sim}{\mu}_i$ para a população Π_i e variância

$V(Y) = \underset{\sim}{c}' Cov(X) \underset{\sim}{c} = \underset{\sim}{c}' \underset{\sim}{\Sigma} \underset{\sim}{c} = \sigma_y^2$ para todas as populações.

Conseqüentemente, o valor esperado global $\bar{\mu}_y$ é definido conforme mostra [3.21] a seguir:

$$\bar{\mu}_y = \frac{\sum_{i=1}^{g} \mu_{iy}}{g} = \underset{\sim}{c}' \underset{\sim}{\bar{\mu}} \qquad [3.21]$$

Assim, a razão entre a soma dos quadrados das distâncias das médias populacionais para a média global e a variância de Y generalizam o caso de duas populações e medem a variabilidade entre grupos de valores Y relativamente à variabilidade conjunta dentro dos

grupos. Logo, selecionamos $\underset{\sim}{c}$ normalizado, de tal forma que $\underset{\sim}{c}'\sum\underset{\sim}{c}=1$, que maximiza a razão conforme se mostra a seguir em [3.22].

$$\frac{\sum_{i=1}^{g}(\underset{\sim}{\mu}_{iY}-\underset{\sim}{\bar{\mu}}_{Y})^{2}}{\sigma_{Y}^{2}}=\frac{\sum_{i=1}^{g}(\underset{\sim}{c}'\underset{\sim}{\mu}_{i}-\underset{\sim}{c}'\underset{\sim}{\bar{\mu}}_{Y})^{2}}{\underset{\sim}{c}'\sum\underset{\sim}{c}}=\frac{\underset{\sim}{c}'B_{0}\,\underset{\sim}{c}}{\underset{\sim}{c}'\sum\underset{\sim}{c}} \qquad [3.22]$$

Seja $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_s > 0,\ \text{onde}\ s \leq \min\{g-1, p\}$ os autovalores não nulos de $\sum^{-1} B_0$ e $\underset{\sim}{e}_1, \underset{\sim}{e}_2, ..., \underset{\sim}{e}_s$ os autovetores correspondentes. O vetor dos coeficientes $\underset{\sim}{c}$ que maximiza a razão $\frac{\underset{\sim}{c}'B_0\,\underset{\sim}{c}}{\underset{\sim}{c}'\sum\underset{\sim}{c}}$ é dada por $\underset{\sim}{c}_1=\underset{\sim}{e}_1$, e a combinação linear $[\underset{\sim}{c}_1'\ \underset{\sim}{X}]$ é chamada de 1ª discriminante de Fisher. Porém, como os parâmetros populacionais são geralmente desconhecidos utiliza-se os resultados amostrais, conforme a seguir:

$$\begin{cases} \underset{\sim}{\bar{x}}_i = \dfrac{1}{n_i}\sum_{i=1}^{g}\underset{\sim}{x}_{ij} \\ \underset{\sim}{\bar{x}} = \dfrac{\sum_{i=1}^{g} n_i\,\underset{\sim}{\bar{x}}_i}{\sum_{i=1}^{g} n_i} = \dfrac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\underset{\sim}{x}_{ij}}{\sum_{i=1}^{g} n_i} \end{cases} \qquad [3.23]$$

A matriz “soma de produtos cruzados entre grupos” é estimada por:

$$\hat{B}_0 = \sum_{i=1}^{g} n_i\,(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})' \qquad [3.24]$$

e o estimador S_p da matriz de covariância $\sum$ pode ser conseguido conforme [3.25].

$$S_p = \frac{W}{\nu} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)'}{n_1 + n_2 + ... + n_g - g} = \frac{\sum_{i=1}^{g} (n_i - 1) S_i}{n_1 + n_2 + ... + n_g - g} \qquad [3.25]$$

Assim, o vetor de coeficientes $\underset{\sim}{\hat{c}}$ que maximiza a razão $\frac{\underset{\sim}{\hat{c}}' \hat{B}_0 \underset{\sim}{\hat{c}}}{\underset{\sim}{\hat{c}}' W \underset{\sim}{\hat{c}}}$ é dada por $\underset{\sim}{\hat{c}}_k = \underset{\sim}{\hat{e}}_k$ e a combinação linear Y é dada em [3.26] por:

$$Y = \underset{\sim}{\hat{c}}'_k \underset{\sim}{X}, \qquad \text{para } k = 1, 2, ..., s \text{ e } k \leq s \qquad [3.26]$$

que é chamada de k-ésimo discriminante.

C. REGRESSÃO LOGÍSTICA PARA CLASSIFICAÇÃO ENTRE DUAS POPULAÇÕES

A regressão logística [Dobson, 1983] consiste em relacionar, por meio de modelos, uma variável resposta *Y* dicotômica com os fatores X_i, *i=1, 2, ..., p-1*, que influenciam as ocorrências de determinado evento. Um exemplo de variável resposta dicotômica poderia ser a rejeição (y=0) ou não rejeição (y=1) de um órgão transplantado [CHAVES Neto, 2001].

Seja o modelo linear logístico simples derivado da função matemática f(y) (sigmoidal), como dado em [3.49]

$$f(y) = \frac{1}{1 + e^{-y}} = \frac{e^y}{1 + e^y} \qquad y \in R, \qquad [3.49]$$

que varia monotonicamente de 0 a 1, à medida que y cresce. Esta função é simétrica em torno de 0,5 e é possível escrever que

$$f(\mathrm{y}) = \mathrm{E}(\mathrm{y} \mid \mathrm{x}) = \frac{e^{y}}{1+e^{y}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad , \qquad [3.50]$$

e a transformação *LOGIT* $\pi(x)$ é dada por:

$$\pi(x) = \ln[f(\mathrm{y})] = \ln[\mathrm{E}(\mathrm{y} \setminus \mathrm{x})] = \ln\left(\frac{f(y)}{1-f(y)}\right) = \ln\left(\frac{e^{\beta_0+\beta_1 x}}{1-e^{\beta_0+\beta_1 x}}\right)$$

$$\pi(x) = \ln\left(\frac{e^{y}}{1-e^{y}}\right) = \ln\left(\frac{1}{1+e^{-y}} \div (1-\frac{1}{1+e^{-y}})\right) = \ln\left(\frac{(1+e^{-y})^{-1}}{e^{-y}(1+e^{-y})}\right) =$$

$$\pi(x) = \ln(1+e^{-y}) - (-\mathrm{y}) - (-\ln(1+e^{-y})) = -\ln(1+e^{-y}) + (\mathrm{y}) + (\ln(1+e^{-y})) = y$$

$$\pi(x) = \mu = \beta_0+\beta_1 x \qquad [3.51]$$

que é o modelo de Regressão Linear Logístico Simples.

A importância desta transformação é que f(x) tem diversas propriedades do modelo de regressão linear [JOHNSON, 1998]:

- A função LOGIT f(x) é linear nos seus parâmetros;
- A regressão linear $y = E(y \mid x) + \varepsilon$, onde *e* tem distribuição normal, $\varepsilon \sim N(0,\sigma^2)$;
- Os modelos têm variáveis dicotômicas;
- ε pode assumir um dos possíveis valores:

$$\begin{cases} Se\ \mathrm{y}=1 \quad \rightarrow \quad \varepsilon = 1-\pi(\mathrm{x}) \\ Se\ \mathrm{y}=0 \quad \rightarrow \quad \varepsilon = -\pi(\mathrm{x}) \end{cases}$$

Portanto o modelo ajustado, que dá as probabilidades de classificação é dado por:

$$\hat{\pi}(\mathrm{x}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} \qquad [3.53]$$

e um novo registro (x_o) é alocado da seguinte maneira:

$$\begin{cases} Se\ \hat{\pi}(x_0) < 0{,}5 \ \rightarrow\ x_0 \text{ é alocado em } \pi_1 \\ Se\ \hat{\pi}(x_0) > 0{,}5 \ \rightarrow\ x_0 \text{ é alocado em } \pi_2. \end{cases}.$$

D. REGRESSÃO LOGÍSTICA ENVOLVENDO MAIS DE UMA VARIÁVEL INDEPENDENTE

Quando o interesse está em se estabelecer uma relação entre a variável resposta *Y* e diversas co-variáveis $x_1, x_2, \ldots, x_{p-1}$, o modelo é chamado de Logístico Linear Múltiplo e tem a seguinte forma:

$$\text{LOGIT } \pi(\underset{\sim}{x}) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} \rightarrow \pi(\underset{\sim}{x}) = \frac{e^{\mu}}{1 + e^{\mu}} \qquad [3.54]$$

$$\text{onde } \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} = \underset{\sim}{x}'\underset{\sim}{\beta}$$

2.7.3. OUTRAS METODOLOGIAS DE CLASSIFICAÇÃO ENTRE DUAS OU MAIS POPULAÇÕES

A. REDES NEURONAIS OU NEURAIS

As Redes Neurais Artificiais (RNA) consistem em um método de solucionar problemas, construindo um sistema que tenha circuitos que simulem o cérebro humano, aprendendo, errando e fazendo descobertas (LEÃO, 2000). Uma grande rede neuronal artificial pode ter centenas ou milhares de unidades de processamento, enquanto os cérebros dos mamíferos possuem cerca de 200 bilhões de neurônios.

De forma geral, a operação de uma célula da rede neuronal artificial pode ser resumida do seguinte modo:

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

Assim como o sistema nervoso dos mamíferos é composto por bilhões de células nervosas, a rede neuronal artificial também é formada por unidades que nada mais são que pequenos módulos que simulam o funcionamento de um neurônio. Estes módulos devem funcionar de acordo com os elementos em que foram inspirados, recebendo e retransmitindo informações. Esta arquitetura divide a rede em neurônios de entrada, que recebem estímulos do meio externo, neurônios internos ou ocultos (*hidden*) e neurônios de saída, que se comunicam com o exterior. As unidades de entrada recebem sinais do meio ambiente, as de saída enviam sinais para o meio ambiente e as escondidas não interagem diretamente com o ambiente (FAUSETT, 1999).

O primeiro *neuro* computador (*Mark I Perceptron*) surgiu em 1957, criado por Frank Rosemblentt. A Rede *Perceptron* tem somente uma camada é pode ser vista como instrumento de reconhecimento de padrões que tem a habilidade de aprender a reconhecer padrões em um conjunto de dados, com convergência após um número finito de iterações, quando a solução existe (STEINER, 1995).

O teorema de Kolmogorov [HECHT-NIELSEN, 1991] afirma que uma rede neuronal com apenas uma camada escondida pode calcular uma função arbitrária qualquer a partir dos dados fornecidos.

Para resolver problemas mais complexos, foi concebido o "*Multilayer Perceptron*", que é uma forma de arranjar *perceptrons* em múltiplas camadas. Para isto, são necessárias mais conexões, os quais só existem em uma rede de *perceptrons* dispostos em camadas. Os neurônios internos são de suma importância na rede neuronal, pois se

provou que, sem estes, torna-se impossível resolver problemas linearmente dependentes (GEMAN, BINENSTOCK E DOURSAT, 1992). Em outras palavras, pode-se dizer que uma rede é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento “inteligente” de uma Rede Neuronal Artificial vem das interações das unidades de processamento da rede.

A maioria dos modelos de redes neuronais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados, ou seja, aprendem por meio de exemplos (KFOURI, 2003). Arquiteturas neuronais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior.

A rede neuronal passa por um processo de treinamento a partir dos casos reais conhecidos, adquirindo, a partir daí, a sistemática necessária para executar adequadamente o processo desejado dos dados fornecidos. Sendo assim, a rede neuronal é capaz de extrair regras básicas a partir de dados reais, diferindo da computação programada, onde é necessário um conjunto de regras rígidas pré-fixadas e algoritmos.

A aplicação de redes neuronais pode ser classificada em classes distintas: Reconhecimento de Padrões e Classificação; Processamento de Imagem e Visão; Identificação de Sistema e Controle e Processamento de Sinais.

A propriedade mais importante das redes neuronais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. Isso é feito por meio de um processo iterativo de ajustes aplicado a seus pesos. O aprendizado ocorre quando a rede neuronal atinge uma solução generalizada para uma classe de problemas (CARVALHO, 1998).

A rede neuronal se baseia nos dados para extrair um modelo geral. Portanto a fase de aprendizado deve ser rigorosa e verdadeira, a fim de se evitarem modelos espúrios. Todo o conhecimento de uma rede neuronal está armazenado nas sinapses, ou seja, nos pesos atribuídos às conexões entre os neurônios. Em torno de oitenta por cento do total dos registros deve ser escolhido aleatoriamente para o treinamento da rede, a fim de que a rede "aprenda" as regras e não "decore" exemplos, permitindo assim a validade do teste. O restante dos dados deve ser utilizado para testes [WITTEN, 1999].

O comportamento de cada unidade de rede pode ser modelado por funções matemáticas simples. A topologia se enquadra no problema em estudo de classificação dicotômica, onde apenas uma unidade de saída é necessária. Uma unidade recebe os sinais de entrada e os agrega com base em uma função de entrada. Esta função de entrada gera um sinal de saída para um determinado padrão, utilizando a função de transferência, que são funções contínuas e diferenciáveis, sugeridas por Rumelhart [RUMELHART, 1986].

O algoritmo *back-propagation* procurará oferecer, por um conjunto de pesos, o melhor ajuste para os pesos apresentados no início do processo. Nesta propagação, os padrões alimentam a rede [RUMELHART, 1986]. O valor de saída obtido para este padrão é comparado com o valor de saída desejado, calculando-se o erro quadrático. O objetivo é minimizar este erro, ajustando o peso de tal modo que todos os vetores de entrada sejam corretamente mapeados em suas correspondentes saídas.

B. ALGORITMO J4.8 – ÁRVORE DE PODA

O algoritmo J4.8 cria uma árvore de decisão construída com poda parcial, com base no algoritmo heurístico C4.5 de Quinlan.

Usa-se este algoritmo quando se quer conhecer qual classe de valores o algoritmo de aprendizagem prediz para cada registro. O

tempo computacional depende da complexidade da árvore gerada (QUINLAN, 1993).

O problema da aprendizagem está na determinação das hipóteses para descrever o conceito alvo que seja consistente com os exemplos de treino e com a teoria sobre o domínio. O algoritmo C4.5 prefere árvores mais simples às mais complexas. Também prefere árvores que colocam os atributos com maior ganho de informação, concentrado junto à raiz das árvores.

Cada nó interno da árvore representa um atributo; cada ramo corresponde a um valor possível para esse atributo; cada folha estabelece uma classificação; cada caminho raiz-folha define uma regra (QUINLAN, 1993).

Para selecionar os melhores atributos com classificador, calcula-se a Entropia (S) [COSTA, 2001], dada por [3.44] a seguir.

$$\text{Entropia(S)} \equiv \sum_{i=1}^{C} -p_i \log_2(p_i) \qquad [3.44]$$

- Se Entropia(S) = 1, máxima desordem
- Se Entropia(S) =0, ausência de desordem

Regra de Quinlan: dado um conjunto S, deve-se escolher o atributo que maximiza o valor do Ganho(S,A). Então, se *v* pertence aos valores(A), o Ganho(S,A) é dado por:

$$\text{Ganho(S,A)} = \text{Entropia(S)} - \sum_{v} \frac{|S_v|}{|S|} \text{Entropia } (S_v), \qquad [3.45]$$

C. ALGORITMO “DECISION STUMP”

Este método de distribuição retorna à distribuição de probabilidades em dado instante. O algoritmo constrói uma árvore de

decisão simples com um nível binário e produz classes de probabilidade [WITTEN, 1999].

D. ALGORITMO IBK – INSTANCE BASED LEARNING

É uma implementação do classificador vizinho-mais-próximo. A opção “validação cruzada” com uma porcentagem dos registros deixados para teste pode ser usada [WITTEN, 1999].

E. ALGORITMO LOGIT BOOST

Este método é baseado no conceito de aditividade da regressão logística. Um método derivado é impulsionado com re-amostras que serão executadas ao invés de reutilizar os pesos [WITTEN, 1999].

F. ALGORITMO ZERO R

O mais primitivo dos algoritmos é chamado de *Zero R*. Ele simplesmente prediz a classe majoritária dos dados treinados.

G. ALGORITMO NAIVE BAYES

Este método tem o nome de *Naive Bayes* por ser baseado em regras condicionais de *Bayes*.

Os acontecimentos para todos os possíveis valores de cada atributo são contados e as probabilidades esperadas são calculadas. Todas as características e as probabilidades globais são tratadas como

intervalos independentes e as frações correspondentes são multiplicadas [WITTEN, 1999], conforme [3.46].

$$\text{Probabilidade do eventos} = \begin{cases} \mathrm{P}(\mathrm{A}_{[\text{sim}]}) = \frac{a_1}{T} * \frac{a_2}{T} * \ldots * \frac{a_n}{T} = p_A \\ \mathrm{P}(\mathrm{B}_{[\text{não}]}) = \frac{b_1}{T} * \frac{b_2}{T} * \ldots * \frac{b_n}{T} = p_B \end{cases} \quad [3.46]$$

Aí, comparam-se as probabilidades p_A e p_B para se saber qual é a mais provável acontecer conforme apresentada a seguir.

$$\begin{cases} \mathrm{P}(\mathrm{A}) = \frac{p_A}{p_A + p_B} \\ \mathrm{P}(\mathrm{B}) = \frac{p_B}{p_A + p_B} \end{cases}, \qquad \text{onde } \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}) = 1 \qquad [3.47]$$

A regra de *Bayes* diz quando se tem uma hipótese (*H*) e evidências (*E*) que afetam as hipóteses. Então, o aprendizado de máquina pesquisa a evolução deste algoritmo, conforme colocado a seguir em [3.48]

$$\mathrm{P}(\mathrm{H} \setminus \mathrm{E}) = \frac{P(E \setminus H) * P(H)}{P(E)} \qquad [3.48]$$

onde *P(E\H)* é a probabilidade do evento *E* acontecer condicionado ao *H*.

Naive Bayes trabalha satisfatoriamente quando testado em conjunto de dados atuais, particularmente quando combinado com procedimentos que servem para eliminar atributos redundantes e dependentes.

2.7.4. AVALIAÇÃO DE FUNÇÃO DE CLASSIFICAÇÃO

O conjunto de dados do treinamento é usado para construir o classificador, enquanto o conjunto de dados para teste é usado para avaliar o desempenho do classificador.

Conforme JOHNSON em 1998, um importante modo de julgar o desempenho de um procedimento de classificação é calcular a Taxa Ótima de Erro de Classificação, OER, que pode ser estimada pela probabilidade total de erro de classificação – TPM, conforme [3.27] a seguir:

$$TPM = p_1 \int_{R_2} f_1(\underset{\sim}{x})\delta\, \underset{\sim}{x} \quad + \quad p_2 \int_{R_1} f_2(\underset{\sim}{x})\delta\, \underset{\sim}{x} = \text{OER} \qquad [3.27]$$

$$\text{onde} \quad \begin{cases} R_1 \;\; \therefore \;\; \dfrac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \dfrac{p_2}{p_1} \\ R_2 \;\; \therefore \;\; \text{em caso contrário} \end{cases} \qquad [3.28]$$

Como normalmente não se tem a forma da distribuição dos dados, uma medida de performance que não depende da forma da distribuição, e que pode ser calculada para qualquer procedimento de classificação, é a Taxa Aparente do Erro - APER, que é definida como a fração das observações no treinamento amostral correspondente ao reconhecimento equivocado pela função. A APER é calculada a partir da matriz de confusão, que mostra a situação real das observações nos grupos versus as observações classificadas (preditas) pelo processo (JOHNSON, R. A., 1998).

Para n_1 observações da população Π_1 e n_2 observações da população Π_1, a matriz de confusão tem a seguinte forma:

$$\begin{array}{lc} & \begin{array}{c}\text{POPULAÇÃO}\\ \text{PREDITA}\end{array} \\ & \begin{array}{ccc}\Pi_1 & \Pi_2 & \text{soma}\end{array} \\ \text{POPULAÇÃO ATUAL} & \begin{array}{c}\Pi_1\\ \Pi_2\end{array}\begin{bmatrix} n_{1C} & n_{1M2} \\ n_{2M1} & n_{2C}\end{bmatrix}\begin{array}{c} n_1 \\ n_2\end{array}\end{array} \qquad [3.29]$$

onde n_{1c} é o número de registros da população Π_1 classificados corretamente como registros da população Π_1, n_{2c} é o número de registros da população Π_2 classificados corretamente como registros da população Π_2, n_{1M2} é o número de registros da população Π_1 classificados incorretamente como registros da população Π_2, n_{2M1} é o número de registros da população Π_2 classificados incorretamente como registros da população Π_1.

Assim, a taxa aparente de erro, que é entendida como a proporção de registros do conjunto de treinamento reconhecidos erroneamente (APER) é dada por [3.30]:

$$APER = \frac{n_{1M2} + n_{2M1}}{n_1 + n_2}\ , \qquad [3.30]$$

No caso de três populações a matriz de confusão tem a forma:

$$\begin{array}{lc} & \begin{array}{c}\text{PREDITA}\end{array} \\ & \begin{array}{cccc}\Pi_1 & \Pi_2 & \Pi_3 & \text{soma}\end{array} \\ \text{ATUAL} & \begin{array}{c}\Pi_1\\ \Pi_2\\ \Pi_3\end{array}\begin{bmatrix} n_{1c} & n_{1m2} & n_{1m3} \\ n_{2m1} & n_{2c} & n_{2m3} \\ n_{3m1} & n_{3m2} & n_{2c}\end{bmatrix}\begin{array}{c} n_1 \\ n_2 \\ n_3\end{array}\end{array} \qquad [3.35]$$

onde os n_{ic}, i=1, 2 e 3, que estão na diagonal principal são os números de registros da população Π_i classificados corretamente como da

população Π_i; n_{imj} são as quantidades de registros da população Π_i classificados incorretamente como da população Π_j, j=1, 2 e 3.

Então, para três populações, a taxa aparente de erro (APER) é dada por [3.32], conforme a seguir:

$$APER = \frac{n_{1m2} + n_{1m3} + n_{2m1} + n_{2m3} + n_{3m1} + n_{3m2}}{n_1 + n_2 + n_3} \qquad [3.32]$$

Uma abordagem que trabalha satisfatoriamente é a abordagem chamada de Lachenbruch ou Validação Cruzada (Lachenbruch & Mickey, 1968). Esta abordagem também é comumente conhecida como "deixando h=1 registros-de-fora" para testes ou "*Leaving-one-out*". É uma técnica que utiliza um estimador não viciado para se estimar os erros de classificação [Duda and Hart, 1973].

Por ser este um método computacionalmente caro, tem sido freqüentemente reservado para problemas onde o tamanho da amostra (n) é relativamente pequeno, pois (n-h) iterações são executadas, deixando de fora h registros, sendo h a proporção dos registros aleatoriamente deixados de fora para teste em cada uma das 1/h iterações. Assim, todos os registros são usados para treinamento e teste, obedecidos os critérios a seguir, conforme JOHNSON, 1998:

Inicia-se com o grupo da população *1*. Deixamos fora *h* registros (*leave-h-out*) para teste. Constrói-se a função discriminante baseada nas (n_1 - *h* + n_2) registros, onde n_1 e n_2 são respectivamente os tamanhos de cada uma das populações. Classificam-se os h registros deixados de fora, utilizando a função construída. Repetem-se os passos *1* e *2* até que todos os registros da população *1* estejam classificados. Seja n_{1m} o número de registros reconhecidos erroneamente no grupo *1*. Repetem-se os passos de *1* a *3* para os n_2 registros da população *2*. Seja n_{2m} o número de registros reconhecidos erroneamente no grupo *2*.

Calculam-se as estimativas das probabilidades P(2\1) e P(1\2) condicionais dos registros erroneamente classificados e a proporção do erro esperado E(AER), como em [3.33]:

$$\hat{P}(2 \setminus 1) = \frac{n_{1M}}{n_1} \qquad \text{e} \qquad \hat{P}(1 \setminus 2) = \frac{n_{2M}}{n_2} \qquad [3.33]$$

e a estimativa da proporção total esperada de erro é dada por:

$$\hat{E}(\text{AER}) = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \qquad [3.34]$$

3. MATERIAL E MÉTODO

3.1. POPULAÇÃO E AMOSTRA

A população é o conjunto de dados dos pacientes doadores e receptores da medula óssea com *Anemia Aplásica Severa* – AAS.

O conjunto dos duzentos registros utilizados neste trabalho foi extraído do Banco de Dados do Serviço de Transplante de Medula Óssea do Hospital de Clínicas da Universidade Federal do Paraná.

3.2. VARIÁVEIS

As variáveis em estudo, em um total de trinta e quatro, são os dados pessoais, laboratoriais e médicos dos pacientes receptores e doadores. Estas variáveis são coletadas em três momentos: antes do paciente ser submetido ao transplante (pré-transplante), conforme tabela 3.1 na página seguinte, e depois que o paciente foi transplantado (pós-TMO), conforme tabela 3.2.

3.3. FERRAMENTAS COMPUTACIONAIS UTILIZADAS

A) WEKA

A escolha do pacote WEKA para realização do trabalho deu-se em virtude de sua facilidade de uso e pela disponibilização de diversos algoritmos que implementam técnicas de mineração de dados.

Atualmente existem sistemas que contêm os métodos de Data mining. Alguns são genéricos e necessitam de maior grau de compreensão do processo por parte dos usuários, mas também existem sistemas projetados para aplicações específicas que exigem pouco conhecimento do processo (GOEBEL, 1999).

TABELA 3.1– GRUPO DE VARIÁVEIS DO PRÉ-TRANSPLANTE DO TMO.

VARIÁVEIS	Descrição	Atributo	Domínio	PRÉ
AboDiferença	Diferença entre tipo sangüíneo do receptor e doador	Categórico	1(diferente) 2 (igual)	Pr1
AboRecep	Tipo sangüíneo e Rh do paciente receptor	Categórico	1=AB 2=A 3=O 4=B	Pr2
CondicCFA	Condicionamento da medula: dose total de ciclofosfamida recebido.	Categórico	1 (cfa) 2 (cfa+bus)	Pr3
DurDoenca	Tempo entre diagnóstico e tmo, em meses	Numérico	Real	Pr4
Etiologia	Etiologia: origem (causas) da doença; Domínio: idiopática ou indeterminado	Categórico	1 2 3	Pr5
HepatitePre	História de hepatite no pré-transplante	Categórico	1=não 2=sim	Pr6
Imunoprofilaxia	Imunoprofilaxia recebido pelo paciente	Categórico	1=não 2=sim	Pr7
ImunossupresPre	Imunossupressão Pré-TMO. Tratou-se a doença	Categórico	1=não 2=sim	Pr8
InfecPreCondic	Paciente teve alguma infecção importante pré-condicionamento	Categórico	1=não 2=sim	Pr9
KarnofLanskyPre	Situação clínica pré-TMO, escala de Karnofsky (IBMTR) Karnofsky (>16 anos) e Lansky (<16 anos)	Numérico	100%=bom	Pr10
NeutrófIntern	Número de neutrófilos no internamento do paciente (valor - hemograma) = (nºleucócitos*neutróf%)/100 ($*10^3$)/µl	Numérico	Real	Pr11
NºCelInfundidas	Número de células do doador infundidas no paciente	Numérico	Real	Pr12
PlaqInt	Plaqueta no internamento do paciente (hemograma)	Numérico	Real	Pr13
SexoDifer	Diferença entre o sexo do Receptor e do Doador	Categórico	1(igual) 2(diferente)	Pr14
transfusãoprévia	Paciente recebeu Transfusões antes do transplante.	Numérico	Real	Pr15
TratamentoPrévio	Se paciente teve tratamento prévio. Normalmente é imunossupressão antes do TMO	Categórico	Exemplos: 0(nada) 1(oximetal) 2(hemogenin) 3(linfoglobu) 4(csa) 5(dexametaz)	Pr16

FONTE: SERVIÇO DE TRANSPLANTE DE MEDULA ÓSSEA DO HC-UFPR

TABELA 3.2– GRUPO DE VARIÁVEIS DO PRÉ E PÓS-TMO.

VARIÁVEIS	Descrição	Atributo	Domínio	Pré/pós
EtniaReceptor	Etnia (raça) do paciente receptor	Categórico	1Bra 2Mul 3Neg 4Ori 5Hispâ6Cauc 7ÍndioAm	Pr17
IdRec-IdDoa	Diferença entre Idade do Receptor e doador	Numérico	Real	Pr18
IdadeReceptor	Idade do paciente na data do Transplante	Numérico	Real	Pr19
SexoReceptor	Gênero do paciente	Categórico	1fem 2masc	Pr20
CistitePós	Paciente teve cistite pós-transplante	Categórico	1=não 2=sim	Po1
Evolução	Paciente vivo/morto	Categórico	1óbito 2vivo	Po2
GrauGVHDA	Grau da doença do enxerto contra o hospedeiro aguda pós	Categórico	0 1 2 3 4	Po3
GrauGVHDC	Grau do GVHD=DECH - Doença do enxerto contra o hospedeiro crônica pós-tmo.	Categórico	0 1 2	Po4
HemorragiaPós	Hemorragia pós-transplante.	Categórico	1=não 2=sim	Po5
HipertensPós	Hipertensão arterial pós-transplante	Categórico	1=não 2=sim	Po6
Imunossupressão	Se o paciente fez imunossupressão-pós para tratar a doença	Categórico	1=não 2=sim	Po7
karnofLanskyPós	Situação clínica do paciente pós-TMO escala IBMTR de.	Numérico	Inteiro múltiplo de dez	Po8
PapaPósTMO	Unidades de papa de hemáceas recebidas pós-tmo.	Numérico	Real	Po9
PlaquetaPósTMO	Unidades de plaquetas recebidas pós-tmo (bolsas com 35/40 ml).	Numérico	Real	Po10
PneumInterstPós	Pneumonite intersticial pós-tmo (pneunomia + grave p/ infecção)	Categórico	0=não 1=sim	Po11
SomaInfecçõesPós	Nº de infecções-Pós (viral, bacteriana, fúngica ou parasitária)	Categórico	0(nada) 1 2 3 4	Po12
Rejeição_0	Se rejeitou ou não.	Categórico	0=não 1=sim	Y1
ClasseRejeição	Tempo até a rejeição. ♦ Classe1 < 100 dias ♦ 101 < Classe2 < 730 ♦ Classe3 > 730 ou não rejeitou.	Categórico	1, 2 ou 3	Y2

FONTE: SERVIÇO DE TRANSPLANTE DE MEDULA ÓSSEA DO HC-UFPR

Neste trabalho foi utilizado o pacote WEKA desenvolvido na Universidade de Waikato na Nova Zelândia (WITTEN & FRANK, 1999). Este pacote é formado por um conjunto de implementações de diversas técnicas de mineração de dados.

O pacote *Weka* está implementado por meio da linguagem de programação *Java* e utiliza-se do paradigma orientado a objetos. O subpacote *Core* possui algumas classes como A*ttribute* (que representam os atributos "nome" e "tipo")*, Instance* (que armazenam os valores dos atributos e os valores das classes) *e Instances* (que contém um conjunto ordenado de instâncias).

O pacote C*lassifiers* armazena as classes mais importantes do pacote *Weka*, onde se encontram implementações de algoritmos para classificações e predições numéricas. Todos os algoritmos de aprendizado são subclasses da classe C*lassifier* e implementam os métodos *Build Classifiers* e *Classify Instance*.

O pacote *Weka* utiliza o padrão *arff* para seus arquivos de entrada, independentemente do algoritmo utilizado. Este padrão de representação de bases de dados garante a independência e a não obrigatoriedade de ordens de precedência entre os registros, bem como a inexistência de inter-relacionamentos entre eles.

A tabela a seguir apresenta o formato de um arquivo com extensão ARFF, isto é, nome do arquivo (@*Relation* Nome), o bloco de definições dos atributos (@*Attribute* Nome-do-Atributo e Domínio do Atributo) que pode ser contínuo, discreto ou nominal e os dados (@*Data*) são separados por vírgula. Um exemplo dos procedimentos para entrada de dados consta a seguir.

EXEMPLO DE ENTRADA DE DADOS

@*relation*	Nome-do-arquivo	DOMÍNIO
@*attribute*	Nome-da-variável	Real
@*attribute*	Nome-da-variável	{1, 2, 3}
@*attribute*	Nome-da-variável	{vivo, óbito}
@*data*	Dados	

B) STATGRAPHICS

O programa *Statgraphics* foi utilizado para o cálculo de sumários estatísticos, gráficos, coeficientes de correlações, regressão múltipla, análise fatorial, modelos de classificação para dois e três grupos e modelos de regressão logística.

C) MATLAB

O programa MATLAB foi utilizado para o cálculo das matrizes de covariância e de correlação, para construir modelos de classificação, para comparação de resultados com outros pacotes e para a construção de programas para implantar as regras construídas.

4. RESULTADOS E DISCUSSÕES

4.1. ANÁLISE DESCRITIVA DOS DADOS

Neste item são descritos o grupo de variáveis e as correlações mais significativas. Em seguida obtêm-se as estimativas de valores faltantes e faz-se a análise fatorial.

A) GRUPO DE VARIÁVEIS

Com a ajuda dos especialistas em transplante, um conjunto de variáveis (atributos) do Banco de Dados do TMO foi selecionado e pré-processado para que pudesse ser utilizado na classificação, conforme a tabela 4.1.

B) SUMÁRIO ESTATÍSTICO

Para averiguar possíveis falhas nos dados ou independência entre atributos, foram calculados sumários estatísticos, utilizando os registros pré-processados do banco de dados do TMO, conforme tabela 4.1 na página seguinte:

TABELA 4.1 – SUMÁRIO ESTATÍSTICO DAS VARIÁVEIS SELECIONADAS.

ATRIBUTOS	ESTATÍSTICAS							
	Média	Erro padrão	Mediana	Moda	Desvio padrão	Assimetria	Mínimo	Máximo
Abo#diferenç	1,62	0,03	2	2	0,49	-0,48	1	2
AboR	2,66	0,05	3	3	0,72	0,14	1	4
CistitePós	1,09	0,02	1	1	0,28	3,00	1	2
CondCfa1	1,66	0,03	2	2	0,47	-0,68	1	2
DurDoen	8,20	1,12	4	2	15,88	5,42	1	144
Etiol1I2S3H	1,16	0,03	1	1	0,41	2,52	1	3
EtniaR	1,68	0,05	2	1	0,74	0,74	1	4
GrauGVHDA	0,49	0,07	0	0	1,06	2,04	0	4
GrauGVHDC	0,23	0,04	0	0	0,59	2,45	0	2
HemorrPós1	0,26	0,03	0	0	0,44	1,13	0	1
HepatPre	1,11	0,02	1	1	0,31	2,60	1	2
HipertPós	1,25	0,03	1	1	0,43	1,19	1	2
idad# (R-D)	-1,07	0,52	-2	-2	7,35	-0,79	-38	18
IdadeR	19,32	0,66	18,5	13	9,28	0,50	2	46
ImunoPre2	1,18	0,03	1	1	0,39	1,68	1	2
Imunoprof	1,99	0,01	2	2	0,12	-8,04	1	2
Imunossupres	1,12	0,02	1	1	0,33	2,36	1	2
InfecPre	1,26	0,03	1	1	0,44	1,10	1	2
KarnofPos	65,75	3,22	100	100	45,47	-0,72	0	100
KarnofPre	81,20	0,83	90	90	11,76	-1,51	40	90
NeutInt	386,9	26,96	256	84	381,24	1,77	7	2072
NumCelInf	3,20	0,08	2,935	3,17	1,14	1,56	1,38	9,16
PapaPos	8,47	0,57	6	6	8,04	2,91	0	49
PlaqPos	57,22	5,28	36,5	20	74,61	5,90	1	815
PneumPós	0,04	0,01	0	0	0,20	4,73	0	1
Sexo#Difer2	1,45	0,04	1	1	0,50	0,20	1	2
SexoR	1,64	0,03	2	2	0,48	-0,59	1	2
somaInfecPós	1,37	0,07	1	1	0,98	0,22	0	4
TransfPre	42,09	4,32	27	51	61,15	6,53	1	675
TratamPreDr	1,52	0,14	0	0	2,04	0,75	0	5

FONTE: O AUTOR

C) CORRELAÇÕES

Na página seguinte, na tabela 4.2, são apresentados os coeficientes de correlação mais significativos entre os pares de variáveis.

TABELA 4.2 - CORRELAÇÕES MAIS SIGNIFICATIVAS ENTRE VARIÁVEIS.

PAR DE VARIÁVEIS			CORRE-LACÃO	PAR DE VARIÁVEIS			CORRE-LAÇAO
PapaPos	X	PlaqPos	**55,1%**	CistitePós	X	GrauGVHDA	**26,7%**
HemorrPós1	X	KarnofPos	**-39,8%**	CondCfa1	X	Imunossupres	**-25,5%**
KarnofPre	X	PapaPos	**-38,0%**	somaInfPós	X	GrauGVHDC	**25,0%**
PapaPos	X	KarnofPos	**-37,8%**	IdadeR	X	idade#	**24,8%**
DurDoença	X	NeutInt	**37,0%**	KarnofPre	X	Sexo#2	**-24,2%**
somaInfPós	X	KarnofPos	**-35,9%**	KarnofPre	X	CondCfa1	**-23,3%**
somaInfPós	X	GrauGVHDA	**34,7%**	IdadeR	X	Imunossupre	**-23,2%**
GrauGVHDA	X	PapaPos	**34,6%**	HemorrPós1	X	PneumPós	**23,2%**
ImunoPre2	X	TransfPre	**34,0%**	HemorrPós1	X	GrauGVHDA	**22,9%**
KarnofPre	X	KarnofPos	**33,9%**	somaInfecPós	X	HipertPós	**22,8%**
CondCfa1	X	TransfPre	**31,9%**	HepatPre	X	Imunoprofil	**-22,6%**
GrauGVHDA	X	KarnofPos	**-31,5%**	EtniaR	X	CondCfa1	**21,9%**
GrauGVHDA	X	GrauGVHDC	**29,9%**	GrauGVHDA	X	PlaqPos	**21,9%**
PneumPós	X	KarnofPos	**-29,6%**	HemorrPós1	X	PlaqPos	**21,6%**
CistitePós	X	KarnofPos	**-29,6%**	DurDoen	X	ImunoPre2	**21,5%**
idade#	X	NumCellnf	**-28,9%**	Abo#2	X	Sexo#2	**-21,4%**
NeutInt	X	InfecPre	**-28,1%**	InfecPre	X	PapaPos	**20,8%**
ImunoPre2	X	HipertPós	**27,8%**	Abo#2	X	DurDoen	**-20,8%**
CistitePós	X	PapaPos	**27,5%**	IdadeR	X	GrauGVHDC	**20,1%**
KarnofPre	X	InfecPre	**-27,4%**				

FONTE: O AUTOR

Nesta tabela pode-se observar algumas correlações importantes:

- Quanto mais *papa* de sangue, maior o número de plaquetas no pós-transplante (r=55,1%);
- Quanto mais *hemorragia* (r=-39,8%) e mais papa de sangue o paciente recebe depois do transplante (-37,8%), pior a condição (*karnofP*ós) do paciente;
- Quanto melhor está o paciente antes do transplante (*karnofP*ré), menos papa de sangue no pós-transplante ele recebe (r=-38%);

- Quanto maior o grau *GVHDA*, mais papa de sangue no pós-transplante o paciente necessita (r=34,6%) e pior é sua condição (*karnofP*ós) no pós-transplante (r=-31,5%);
- Quanto melhor a condição (*karnofPré*) do paciente no pré-transplante, melhor também é sua condição no pós-transplante (33,9%).

4.2. ESTIMATIVA DE VALORES FALTANTES

Para se estimar alguns valores faltantes do banco de dados, construiu-se modelos de regressão múltipla, utilizando as variáveis independentes com melhor significância. Por exemplo, os valores perdidos ou desconhecidos da variável “Duração da doença” (tempo decorrido entre o diagnóstico da doença e o transplante) foram estimados pela equação [4.1] dada a seguir:

$$Y = 16,8809 - 10,5605X_1 + 4,28273X_2 \qquad [4.1]$$

Esta sentença descreve a relação entre a variável dependente “Duração da Doença” (Y) e as variáveis independentes “Evolução” (X_1) (p=0,018) e “Grupo-de-Rejeição”(X_2) (p=0,076). O quadro 4.1 da *ANOVA* mostra a significância do modelo ajustado, na qual pode-se afirmar que existe relacionamento estatisticamente significativo (p=0,060) entre as variáveis com 94% de nível de confiança.

QUADRO 4.1 – ANOVA – ANÁLISE DA VARIÂNCIA

Fonte	Soma dos Quadrados	Graus de liberdade	Quadrado Médio	F	p-valor
Modelo	2848,91	2	1424,45	2,83	0,060
Residual	118651	236	502,76		
Total	151500	238			

4.3. AGRUPAMENTO DE VARIÁVEIS

A Análise Fatorial mostrou que as variáveis foram agrupadas em doze fatores comuns (autovalores maiores que um que compõem 66,17% da variação explicada), conforme mostra a tabela número 4.3. Na tabela 4.4, pode-se observar que o fator um, por exemplo, agrupou quatro atributos, enquanto o fator doze somente um. Nos fatores de um a quatro somente atributos do pré-transplante foram agrupados. O fator cinco agrupou variáveis tanto do pré quanto do pós-transplante, e os fatores de seis a doze agruparam variáveis que são consideradas do pós-transplante.

TABELA 4.3 –AUTOVALORES, VARIAÇÃO PERCENTUAL E VARIAÇÃO PERCENTUAL ACUMULADA.

Nº	autovalores	Variação%	Var.acumulada%
1	5,11	15,02	15,02
2	2,89	8,49	23,51
3	2,26	6,65	30,16
4	1,87	5,51	35,67
5	1,69	4,98	40,65
6	1,51	4,43	45,09
7	1,37	4,02	49,10
8	1,32	3,88	52,98
9	1,23	3,61	56,59
10	1,13	3,33	59,92
11	1,10	3,25	63,17
12	1,02	3,00	66,17
13	0,97	2,85	69,02
14	0,96	2,82	71,83
15	0,87	2,55	74,38
16	0,82	2,42	76,80
...	...	...	...

QUADRO 4.2 – COMUNALIDADES – PRÉ-TMO

ATRIBUTOS	VALOR	ATRIBUTOS	VALOR
Idade do receptor	0,701946	Condicionamento	0,535601
Neutrófilos na Internação	0,663089	Idade_diferença	0,506342
Abo Receptor	0,660382	Karnof_Pre	0,495281
Duração da Doenca	0,648228	Transfusão	0,486370
Imuno_Pre	0,643905	Hepatite Pre	0,483629
Etiologia_1I2S3H	0,637854	Sexo #	0,474579
Tratamento Pre	0,618932	Abo diferença	0,452135
Etnia Receptor	0,596872	Sexo Receptor	0,322603
InfeccaoPreCondicionada	0,571316		

TABELA 4.4 –ATRIBUTOS DISTRIBUÍDOS PELOS FATORES.

Fator 1	NeutrófInternam	Infecção Pré	Karnof Pré	Dur Doença
Fator 2	Etnia do Receptor	Tipo Condic Pré	Hepatite Pré	
Fator 3	Sexo Receptor	Sexo Diferença		
Fator 4	Etiologia	ABO Diferença		
Fator 5	Imunossupres Pré	Tratamento Pré	Transfusão Pré	Hipert Pós
Fator 6	KarnofLansky Pós	Hemorragia Pós	Pneum Int Pós	Evolução
Fator 7	Imunossupres Pós	Idade Receptor	Rejeição	
Fator 8	Idade Diferença	Nº Cel Infundidas		
Fator 9	Plaqueta Pós	Papa Sangue Pós		
Fator 10	ABO Receptor	Imunoprofilax		
Fator 11	Grau GVHD C	Grau GVHD A	Soma Infecções	
Fator 12	Cistite Pós			

FONTE: O AUTOR

QUADRO 4.3 – COMUNALIDADES –PRÉ E PÓS-TMO.

ATRIBUTOS	VALOR	ATRIBUTOS	VALOR
Papa de sangue Pos (1º)	0,773641	Infecção PreCondic	0,615782
soma infecções Pós (2º)	0,745538	Idade receptor	0,612837
GVHDC (3º)	0,731769	Imunos Pos	0,609219
Idade diferença (4º)	0,685697	Imunoprofil	0,609151
Transfusão Previa (5º)	0,666369	Abo#	0,608235
Neutrófilos na Internação	0,661533	Etnia Receptor	0,589114
Karnof Pos (7º)	0,657100	AboR	0,580512
PneumInt Pós (8º)	0,653568	KarnofPre	0,577947
TratamPre_Dra (9º)	0,642864	GVHDA	0,575237
Hipertensão Pós (10º)	0,640692	HemorragiaPós	0,564570
Plaquetas Pos (11º)	0,634615	Condicion_cfa1	0,561452
Células Infundidas (12º)	0,629599	Etiologia 1I2S3H	0,552750
Imuno Pré	0,626774	SexoR	0,533797
Duração da Doenca	0,624061	CistitePós	0,509496
Hepatite Pre	0,616259	Sexo#2	0,506847

4.4. CLASSIFICAÇÃO DE PACIENTES DO TMO

Os especialistas em transplante sempre estão à procura de ferramentas e técnicas que possam auxiliá-los nas decisões quanto ao melhor caminho a ser tomado para evitar a rejeição ou, pelo menos, prolongar o tempo de vida do transplante.

Para ajudar a suprir essas necessidades, este trabalho propõe a construção de regras estatísticas e de aprendizagem de máquina para classificar novos pacientes em classes de rejeição, utilizando o Banco de Dados do Serviço de Transplante de Medula Óssea coletados de pacientes receptores e doadores.

Duas estratégias foram tomadas quanto à classificação dos pacientes:

A) Classificação em duas populações:

- População 0 - pacientes que rejeitaram o TMO
- População 1 - pacientes que não rejeitaram o TMO

B) Classificação em três populações quanto ao tempo até a rejeição:

- População 1 - pacientes que rejeitaram o TMO em menos de 100 dias
- População 2 - pacientes que rejeitaram entre 100 dias e dois anos
- População 3 - pacientes que rejeitaram após dois anos ou que não rejeitaram

4.4.1. CLASSIFICAÇÃO EM DUAS POPULAÇÕES

Os especialistas têm grande interesse em saber de antemão as possibilidades de um paciente submetido ao transplante vir a rejeitar

(população 0) ou não rejeitar (população 1) a medula transplantada. Diversos algoritmos serão utilizados aqui para classificar os pacientes.

4.4.1.1. CLASSIFICAÇÃO DE FISHER PARA DUAS POPULAÇÕES

Passos seguidos na Análise Discriminante:

- SUMÁRIO ESTATÍSTICO DOS DADOS POR GRUPOS POPULACIONAIS
- VETORES DAS MÉDIAS E DOS DESVIOS PADRÕES
- MATRIZES DE COVARIÂNCIAS E CORRELAÇÕES
- MATRIZES DAS VARIAÇÕES
- AUTOVALORES
- FUNÇÃO DE CLASSIFICAÇÃO
- ESPAÇO DISCRIMINANTE
- CENTRÓIDES
- AVALIAÇÃO DO MODELO DE CLASSIFICAÇÃO
- MATRIZ DE CONFUSÃO E PROBABILIDADE A PRIORI
- ESTIMATIVAS – ALOCAR NOVOS REGISTROS

Para comparação entre as populações, foram calculados os sumários estatísticos para cada população (pacientes que rejeitaram e pacientes que não rejeitaram o TMO), utilizando os registros pré-processados do banco de dados do TMO, conforme tabela a seguir:

TABELA 4.5 – SUMÁRIO ESTATÍSTICO DAS VARIÁVEIS ESTRATIFICADAS EM DUAS POPULAÇÕES

	Médias		Medianas		Modas		Erros padrões	
POPULAÇÕES	**Popul 0**	**Popul 1**	**Popul 0**	**Popul 1**	**Pop 0**	**Pop 1**	**Popul 0**	**Popul 1**
Abo#Difer2	1,58	1,63	2	2	2	2	0,06	0,04
AboR	2,74	2,61	3	3	3	2	0,08	0,07
CistitePós	1,20	1,03	1	1	1	1	0,05	0,01
Cond_cfa1	1,77	1,60	2	2	2	2	0,05	0,04
DurDoenca	10,45	7,08	5	3	3	2	2,09	1,32
Etiol 1I2S3H	1,18	1,15	1	1	1	1	0,06	0,03
EtniaRec	1,64	1,70	1	2	1	1	0,10	0,06
GrauGVHDA	0,97	0,25	0	0	0	0	0,17	0,06
GrauGVHDC	0,27	0,20	0	0	0	0	0,09	0,05
HemorPós_1	0,53	0,12	1	0	1	0	0,06	0,03
HepatitePre	1,14	1,09	1	1	1	1	0,04	0,02
HipertPós	1,20	1,27	1	1	1	1	0,05	0,04
IdadeR	20,29	18,84	19	18	14	13	1,15	0,80
idR_idD	-0,97	-1,11	-2	-1	-2	-3	0,90	0,64
ImunoPre_2	1,26	1,14	1	1	1	1	0,05	0,03
Imunoprofil	1,97	1,99	2	2	2	2	0,02	0,01
Imunoss2	1,09	1,13	1	1	1	1	0,04	0,03
InfeccaoPre	1,39	1,19	1	1	1	1	0,06	0,03
KarnofPos	5,61	95,37	0	100	0	100	2,74	1,13
KarnofPre	75,76	83,88	80	90	90	90	1,78	0,78
NeutInt	385,76	387,57	252	275	252	84	49,29	32,22
NumCelInf	3,32	3,13	3,08	2,91	3,17	3,30	0,16	0,09
PapaPos	12,52	6,48	9	6	4	3	1,44	0,36
PlaqPos	76,23	47,85	47	35	15	20	13,74	3,84
PneumIntPós	0,12	0,00	0	0	0	0	0,04	0,00
Sexo#Difer2	1,50	1,43	1,5	1	2	1	0,06	0,04
SexoR	1,61	1,66	2	2	2	2	0,06	0,04
somaInfecPós	1,82	1,14	2	1	2	1	0,11	0,08
TransfPrevia	50,85	37,78	35	23	51	14	6,56	5,57
TratamPre_Dr	1,77	1,39	0	0	0	0	0,26	0,17

FONTE: O AUTOR

Utilizando o banco de dados do TMO com registros pré-processados, funções de classificação foram construídas para alocar novos pacientes em uma das duas populações (rejeitar ou não rejeitar).

Utilizando todas as variáveis do pré-transplante para classificação, a taxa de acerto ficou em 83% (ver anexo 1). Porém, depois de várias simulações de variáveis, e com a ajuda dos especialistas em transplante, encontrou-se um grupo de variáveis do pré-transplante com somente quatro variáveis (duração da doença, infecção, karnofPré e tipo de condicionamento pré), com taxa de acerto de 80,5%. Portanto, este grupo com quatro variáveis foi utilizado para construir as funções de classificação e alocar novos pacientes em uma das duas populações.

Os vetores médios das quatro variáveis para cada população e global estão descritos na tabela a seguir:

TABELA 4.6 - VETORES DAS MÉDIAS DOS QUATRO ATRIBUTOS

Atributos	Média da População 0	Média da População 1	Média Global
Duração da Doença	10,692	7,317	8,195
Infecção pré	1,115	1,311	1,260
Karnof Pré	84,808	79,932	81,200
Tipo de Condicionamento	1,327	1,777	1,660

FONTE: O AUTOR

NOTA: 0 = População dos que rejeitaram o TMO; 1= População dos que não rejeitaram.

As matrizes de covariâncias apresentadas a seguir serão utilizadas na Análise de Classificação.

MATRIZ DE COVARIÂNCIAS DA POPULAÇÃO 0

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	595	-0,67	-17,51	1,65
InfecçãoPré		0,10	-0,57	0,06
KarnofPré			45,06	-1,41
Condicionam				0,22

MATRIZ DE COVARIÂNCIAS DA POPULAÇÃO 1.

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	132	-0,59	-28,62	0,75
InfecçãoPré		0,22	-1,48	0,009
KarnofPré			165,3	-0,70
Condicionam				0,22

MATRIZ DE COVARIÂNCIA CONJUNTA

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	251,3	-0,61	-25,8	0,98
InfecçãoPré		0,19	-1,24	0,02
KarnofPré			134,3	-0,88
Condicionam				0,19

MATRIZ DE CORRELAÇÃO CONJUNTA

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	1	-0,09	-0,14	0,14
InfecçãoPré		1	-0,25	0,12
KarnofPré			1	-0,18
Condicionam				1

Esta matriz mostra as correlações conjuntas calculadas entre as variáveis independentes dentro de cada grupo.

As matrizes das variações serão utilizadas na Análise de Classificação, conforme a seguir:

MATRIZ W DAS VARIAÇÕES DENTRO DOS GRUPOS (1E4)

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	4,95	-0,012	-0,510	0,0195
InfecçãoPré		0,004	-0,025	0,0004
KarnofPré			2,660	-0,0174
Condicionam				0,0037

MATRIZ B_O DAS VARIAÇÕES ENTRE OS GRUPOS

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	7,006	-0,406	10,12	-0,935
InfecçãoPré		0,024	-0,586	0,054
KarnofPré			14,622	-1,350
Condicionam				0,125

A função de classificação, calculada a partir dos autovetores, é usada para discriminar novos pacientes nas duas diferentes populações.

Esta função aprendida é usada para discriminar novos pacientes em um dos dois grupos.

- Y = -2,417 – 0,02*DURDOENÇA + 0,483*INFECCAOPRE – 0,0173*KARNOFPRE + 2,04*CONDICIONAMENTO

[4.2]

O modelo depois de padronizado fica:

- Y = -0,32*DURDOENC + 0,21*INFECPRE - 0.20*KARNOFPRE + 0,881*CONDIC

[4.3]

O espaço discriminante nas duas dimensões é obtido substituindo-se os dados originais de todos os registros nas funções discriminantes; assim se obtêm os vetores no espaço discriminante, conforme se mostrado na tabela 4.8, a seguir.

Utilizando as funções aprendidas, os pacientes realocados, conforme mostra a tabela 4.8 na página seguinte. Por exemplo:

- Os pacientes números um e dois pertenciam originalmente à população um (não rejeição) e foram classificados corretamente na população um (não rejeição).

- Os pacientes números três e cinco pertenciam à população zero (rejeição) e foram classificados incorretamente na população um (não rejeição).

O asterisco ao lado do número significa classificação incorreta. Contados os registros sem asterisco, 80,5% dos pacientes foram classificados corretamente.

Centróides são as médias dos espaços discriminantes das duas populações, conforme podemos observar no quadro 4.4 a seguir:

QUADRO 4.4 – CENTRÓIDES

POPULAÇÕES	CENTRÓIDE
0	-0,860402
1	0,302303

TABELA 4.7 - DESVIOS PADRÕES DOS QUATRO ATRIBUTOS PARA AS DUAS POPULAÇÕES E GLOBAL.

Atributos	Desvio padrão da População 0	Desvio padrão da População 1	Desvio padrão Global
Duração da Doença	24,393	11,492	15,882
Com ou sem infecções pré	0,323	0,464	0,440
Karnof Pré (Condição)	6,713	12,857	11,758
Tipo de Condicionamento	0,474	0,418	0,475

FONTE: O AUTOR

TABELA 4.8 - VETORES NO ESPAÇO DISCRIMINANTE (PARTE)

Número Registro	Grupo atual	Grupo predito	Valor F1	Estimador 2	Valor F2
1	1	1	40,71	1	38,52
2	1	1	48,16	2	46,05
3	0	1*	48,62	2	46,63
4	1	1	66,94	2	64,39
5	0	1*	59,87	1	57,32
6	1	1	32,63	2	32,04
7	0	0	45,70	2	45,12
8	0	0	45,01	2	44,57
9	1	1	20,92	3	20,78
10	1	1	4,72	3	4,08
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
197			67,59	2	65,20
198			56,72	2	54,98
199			57,09	2	55,45
200			50,76	1	49,31

FONTE:O AUTOR
NOTA: * CLASSIFICAÇÃO INCORRETA
NOTA: 0 = REJEIÇÃO; 1= NÃO REJEIÇÃO

Para construir a matriz confusão, os dados dos registros atuais foram substituídos nas funções discriminantes e alocados a uma população predita. Quando as populações atuais coincidem com as dos preditos, conta-se como acerto e são anotados na diagonal principal da matriz de confusão, conforme quadro 4.5 a seguir:

QUADRO 4.5 – MATRIZ DE CONFUSÃO EM PERCENTAGEM.

	GRUPO PREDITO (%)		
GRUPO ATUAL	0	1	total
0	67,31%	32,69%	52
1	14,86%	85,14%	148

Porcentagem total de casos corretamente classificados: 80,50%

Pode-se observar na linha zero da matriz confusão que 67,31% dos pacientes que pertenciam ao grupo zero (rejeição) foram classificados corretamente como do grupo zero, enquanto 32,69% pertenciam ao grupo zero e foram classificados como do grupo um. Na última linha, 85,14% que pertenciam à população um (não rejeição) foram classificados corretamente como da população um, enquanto 14,86% que pertenciam à população um, foram classificados incorretamente como da população zero.

As freqüências relativas simples das populações originais, que mostram como as populações a priori se distribuem, estão anotadas no quadro 4.6 a seguir:

QUADRO 4.6 – FREQÜÊNCIAS RELATIVAS

POPULAÇÕES	FREQÜÊNCIAS
0 (Rejeição)	26%
1 (Não Rejeição)	74%

Para classificar um novo indivíduo com estas regras aprendidas, - seus atributos são substituídos nas funções discriminantes, conforme [4.3]. Obtidos os dois valores estimados, calculam-se as distâncias em relação aos centróides. Na população cuja distância for menor, aloca-se o novo indivíduo, conforme colocado em [3.17].

Para o cálculo da taxa de acerto foi utilizado o método da Validação Cruzada, também chamado de Lachembruch ou *Leave-ten-out*, deixando em cada uma das dez rodadas 10% dos registros para teste e os 90% restantes para o treinamento. Os resultados podem ser observados na tabela 4.9, onde AER é a Razão Atual do Erro ou *Actual Error Rate.* No geral, a taxa de acerto ficou em 80,5%.

Outra maneira de observarmos o desempenho do modelo é pelo Custo Esperado de Reconhecimento (ECM) ou Taxa Aparente do Erro, que é dado pela soma dos produtos dos elementos fora da diagonal principal da matriz de confusão.

Observando a tabela 4.9 da Análise Discriminante, na primeira simulação, 66% e 85% respectivamente foram classificados corretamente nos grupos zero ou um conforme mostra a diagonal principal da matriz de confusão. Na última simulação (20ª), essas porcentagens passam para 65% e 84%. No geral, 80,5% foram classificados corretamente.

GRÁFICO 4.1- DIAGRAMA DOS AUTOVALORES

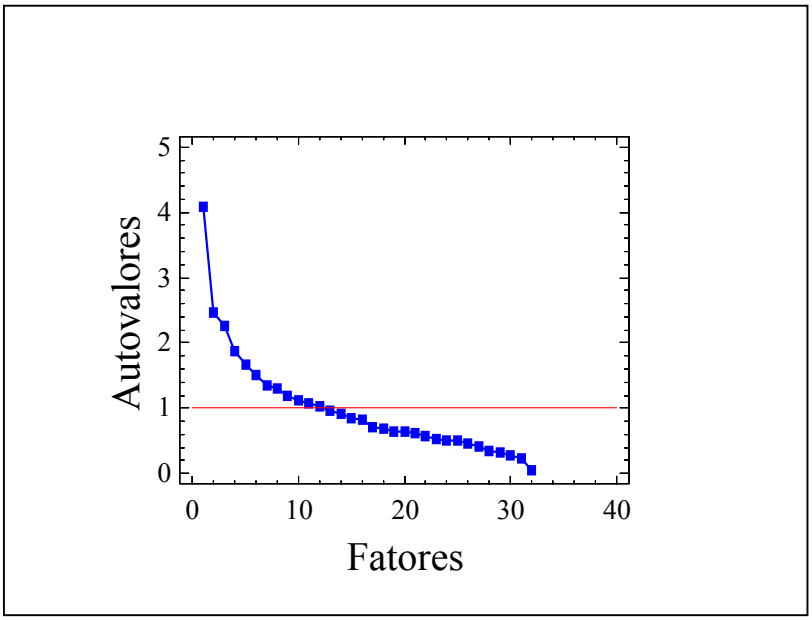


GRÁFICO 4.2 - COMPARATIVO ENTRE AS POPULAÇÕES

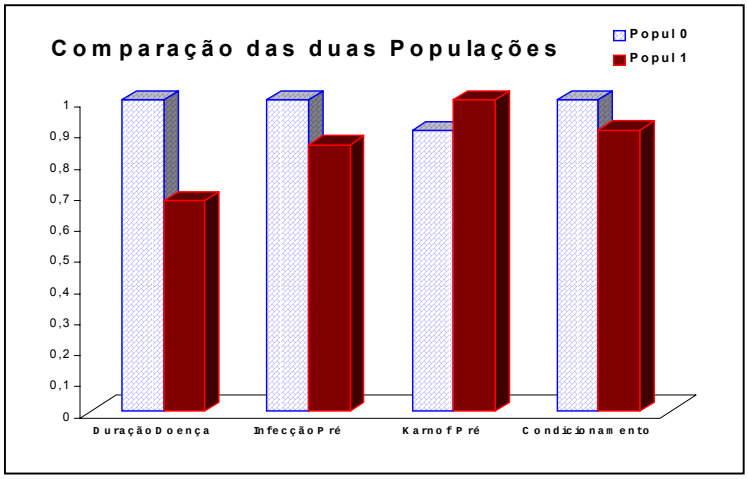


TABELA 4.9 – DEZ REGRAS APRENDIDAS COM 90% DOS REGISTROS.

Coeficientes das funções discriminantes	Matriz de confusão			% dos casos classificados corretamente
	Predito	Atual		
1-20 PARA TESTE		0	1	
DurDoenc -0,313298 InfeccaoPre 0,179345 KarnofPre -0,151394 Cond_cfa1 0,909711	0 1	66% 15%	34% 85%	AER=75% ECM=80%
21-40 PARA TESTE				
DurDoenc -0,243524 InfeccaoPre 0,267058 KarnofPre -0,196354 Cond_cfa1 0,887477	0 1	65% 13%	35% 87%	AER=60% ECM=81%
41-60 PARA TESTE				
DurDoenc -0,387105 InfeccaoPre 0,208381 KarnofPre -0,242942 Cond_cfa1 0,845739	0 1	50% 14%	50% 85%	AER=85% ECM=77%
61-80 PARA TESTE				
DurDoenc -0,190626 InfeccaoPre 0,21627 KarnofPre-0,109673 Cond_cfa1 0,916065	0 1	66% 17%	34% 83%	AER=80% ECM=79%
81-100 P/ TESTE				
DurDoenc -0,361399 InfeccaoPre 0,204968 KarnofPre-0,172953 Cond_cfa10,887971	0 1	67% 15%	33% 85%	AER=90% ECM=81%
101-120 P/ TESTE				
DurDoenc -0,321027 InfeccaoPre 0,226658 KarnofPre -0,228234 Cond_cfa1 0,876274	0 1	69% 15%	31% 85%	AER=80% ECM=81%
121-140 P/ TESTE				
DurDoenc-0,387154 InfeccaoPre0,221484 KarnofPre-0,274104 Cond_cfa1 0,844836	0 1	59% 13%	41% 87%	AER=70% ECM=80%
141-160 P/ TESTE				
DurDoenc-0,327575 InfeccaoPre 0,116612 KarnofPre-0,151948 Cond_cfa1 0,925787	0 1	73% 19%	27% 81%	AER=70% ECM=79%
161-180 P/ TESTE				
DurDoenc -0,315001 InfeccaoPre0,206449 KarnofPre -0,258174 Cond_cfa1 0,845579	0 1	63% 14%	37% 86%	AER=80% ECM=79%
181-200 P/ TESTE				
DurDoenc-0,342342 InfeccaoPre0,237277 KarnofPre -0,207751 Cond_cfa1 0,852203	0 1	65% 16%	35% 84%	AER=90% ECM=79%
TOTAL	0 1	65% 18%	35% 82%	AER=78,0% ECM=80,5%

FONTE: O AUTOR
NOTA: AER = *Actual Error Rate.*

4.4.1.2. REGRESSÃO LOGÍSTICA PARA CLASSIFICAÇÃO ENTRE DUAS POPULAÇÕES

A análise da variância mostra os resultados do ajuste do modelo da regressão logística para descrever a relação entre a variável dependente e as variáveis independentes.

QUADRO 4.7 - ANÁLISE DA VARIÂNCIA - ANOVA

FONTE	VALOR	Graus de Liberdade	p-valor
Modelo	42,2763	4	0,0000
Resíduo	186,9470	195	0,6480
Total	229,2233	199	

Como o p-valor do modelo na ANOVA é menor que 0,0001, podemos afirmar que existe relação estatística significante entre as variáveis no nível de significância de 99,9%. Ainda, como o p-valor (p=0,648) dos resíduos é maior que 10%, indica que este modelo é significativamente aceito no nível de 90% de confiança.

QUADRO 4.8 - COEFICIENTES DO MODELO DE REGRESSÃO

Parâmetros	Estimativa	Erro estimado	razão	p-valor
Dur_Doença	-0,0198	0,0108	0,9803	p=0,058
InfeccaoPré	0,7193	0,5169	2,0529	p=0,140
KarnofPré	-0,0227	0,0192	0,9775	p=0,219
Condicionamemto	0,8951	0,3778	6,6534	p=0,000
Constante	-0,7248	1,9995		

Observa-se no quadro dos coeficientes, que a variável mais significativa é o tipo de condicionamento (p<0,001), seguido pela duração da doença (p<0,06).

Portanto, conforme estimativas mostradas no quadro dos coeficientes, a equação do modelo de ajuste, com taxa de acerto de 77%, é dada por:

$$REJEIÇÃO = e^{\left(\frac{eta}{1+e^{eta}}\right)} \qquad [4.4]$$

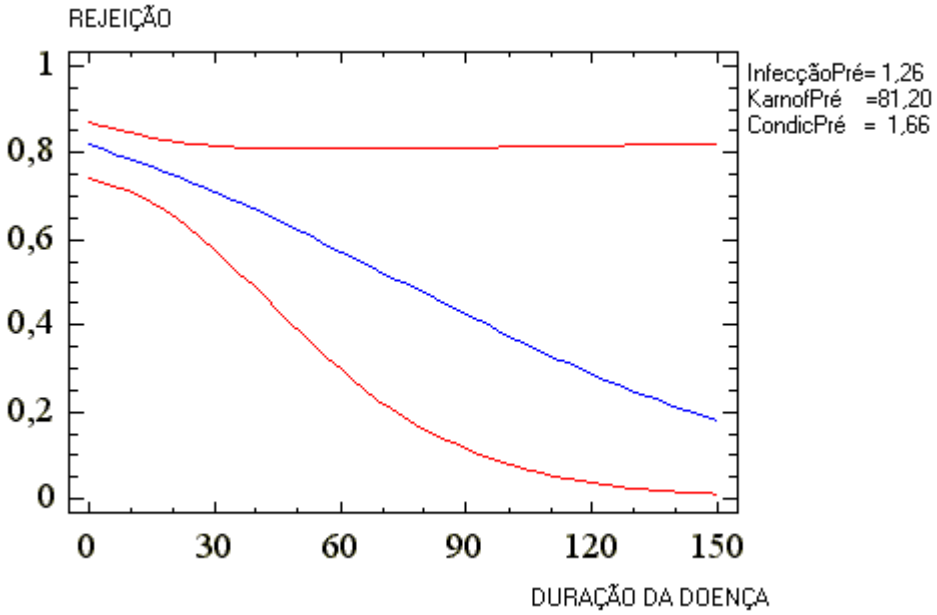
onde

$$eta = -0{,}72 - 0{,}02 * \text{DurDoenc} + 0{,}72 * \text{InfecPre} - 0{,}023 * \text{KarnofPre} + 1{,}9 * \text{CondCfa}$$

Na página seguinte mostram-se alguns diagramas que representam o comportamento da rejeição em função das variáveis.

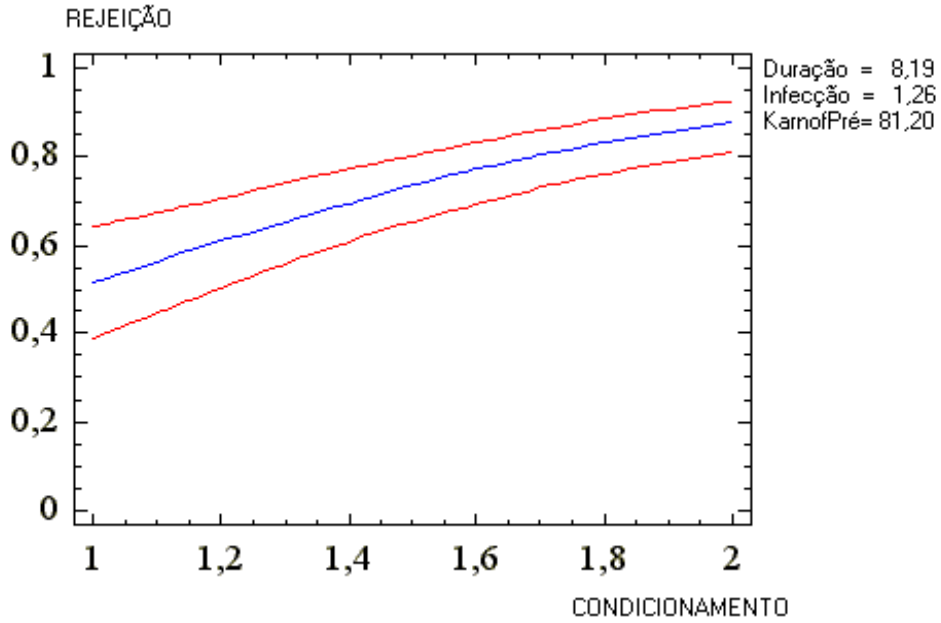
Pode-se observar no diagrama ajustado (gráfico 4.3) que, quanto maior a duração da doença, maior a probabilidade de rejeitar o TMO, e, no segundo diagrama ajustado, a indicação ao paciente do condicionamento tipo 2 (cfa+bussolfan) tem maior probabilidade de não rejeitar.

GRÁFICO 4.3 – DIAGRAMA DA REGRESSÃO LOGÍSTICA – DURAÇÃO DA DOENÇA



FONTE: O AUTOR
NOTA: População 0 = rejeição; população 1 = não rejeição.

GRÁFICO 4.4 – DIAGRAMA DA REGRESSÃO LOGÍSTICA – CONDICIONAMENTO.



FONTE: O AUTOR
NOTA: População 0 = rejeição; população 1 = não rejeição.

4.4.1.3. OUTROS ALGORITMOS UTILIZADOS

Na busca de melhores taxas e novas regras de aprendizagem, lançou-se mão de outros algoritmos de classificação. O pacote utilizado foi o *WEKA*. Os classificadores utilizados se encontram na tabela a seguir, que mostra o nome do algoritmo, a taxa de acerto (Lachenbruch) e a matriz de confusão.

TABELA 4.10 – PERFORMANCE DE ALGUNS ALGORITMOS UTILIZADOS

Algoritmos	Taxa de acerto	Matriz de confusão		
		Atuais	Preditos	
			0	1
Kernel Estimador	79,5%	0	65,4%	34,6%
		1	15,5%	84,5%
Redes Neuronais	78,5%	0	61,5%	38,5%
		1	15,5%	84,5%
J4.8 (C4.5-Quinlan)	77%	0	55,2%	44,8%
		1	20,3%	79,7%
Regressão Logística	77%	0	59,2%	40,8%
		1	15,5%	84,5%
K-Star (B20Mn)	76,5%	0	51,9%	48,1%
		1	14,9%	85,1%
Naïve Bayes Class	76,5%	0	55,8%	44,2%
		1	16,2%	83,8%
Logit Boost	77,5%	0	55,8%	44,2%
		1	14,9%	85,1%
AD Tree (B10 E3)	78%	0	65,4%	34,6%
		1	17,6%	82,4%
IB1 Classifier	74%	0	40,4%	59,6%
		1	14,2%	85,8%
B6 – One-R	69%	0	34,6%	65,4%
		1	18,9%	81,1%
SMO	69%	0	53,8%	46,2%
		1	16,2%	83,8%
Voting Feature Bias	75,5%	0	65,4%	34,6%
		1	20,9%	79,1%
Voted Perceptron 84	73,5%	0	0,0%	100,0%
		1	0,7%	99,3%
Ada Boost (M1)	74%	0	0,0%	100,0%
		1	0,0%	100,0%
Decision Stump	72%	0	50,0%	50,0%
		1	20,3%	79,7%

FONTE: O AUTOR

NOTA: 0 = População dos que rejeitaram o TMO; 1= População dos que não rejeitaram.

DESCRIÇÃO DE ALGUNS ALGORITMOS

A) REDES NEURONAIS OU NEURAIS

Os parâmetros utilizados aqui são: L=0.3, M=0.2, N=1550, V=0, S=0, E=20, NA= normalizado, onde L é razão de aprendizado ou *learning rate*, M são os impulsos ou *momentum*, N é o número de interações, V é a validação ou *validation set size*, S é a semente aleatória, E é a validação ou *validation threshold* e NA mostra se os atributos são ou não normalizados.

O conjunto de treinamento é mostrado a seguir:

CONJUNTO DE TREINAMENTO

ENTRADAS	PESOS
Threshold (Nó Zero)	1,5973939
Nó 2	-1,1353421
Nó 3	-1,1345031
Nó 4	-1,5230746
. . . .	. . . .
Threshold (Último nó)	1,5662866
KarnofLansky Pré	-1,6121413
Duração Doença	-3,0170405
Infecções Pré	2,3849835
Condicionamento	3,106165

CONCLUSÕES

O número de registros classificados corretamente foi de 157 e incorretamente foi de 43 (21.5%), com erro médio absoluto de 0.2895.

Podemos observar que a taxa de acerto pelo aprendizado em redes Neuronais (78,5%) ficou praticamente idêntica à taxa de acerto

pela análise Discriminante (78%). Assim, podemos optar por um dos modelos para inferir novos registros quanto à rejeição ou não do TMO.

B) C4.5 DE QUINLAN - ÁRVORE DE PODA

Os parâmetros estabelecidos em C4.5 são: C = 0,4; M = 2, onde C é o fator de confiança e M o número de quebras. A árvore gerada por este método é dada a seguir.

ÁRVORE - CONJUNTO DE TREINAMENTO [4.5]

Condicionam = 1
| *InfecPre = 1:* ***0***
| *InfecPre = 2:* ***1***
Condicionam = 2
| *KarnofPre <= 80*
| | *DurDoenc <= 1:* ***0***
| | *DurDoenc > 1:* ***1***
| *KarnofPre > 80:* ***1***

CONCLUSÕES DO ALGORITMO C4.5

Esta árvore, de tamanho oito, gastou computacionalmente 1,01 segundo para “aprender” e testar os 200 registros, de vinte em vinte. De todos os registros, 77% foram classificados corretamente. Dos registros da classe zero, 55,2% foram classificados corretamente, enquanto 79,7% dos da classe um (ver tabela 4.10). Observando o conjunto de treinamento, podemos tirar as seguintes conclusões:

- Utilizando o tipo de condicionamento um (cfa) e com infecção pré-transplante, o paciente será alocado à população zero (rejeitar).
- Se não teve infecção, então será alocado à população um (não rejeitar).
- Utilizando o tipo de condicionamento dois (*cfa+bussofan*) e com condição *KarnofPré* menor ou igual a 80, e ainda com duração da

doença menor ou igual que um mês, o paciente será alocado à população zero (rejeitar).

- Porém, se a duração da doença for maior que um mês, então é alocado à população um (não rejeitar) e, se o *karnofPré* for maior que 80, independente da duração da doença, ele é alocado à população um (não rejeitar).

4.4.2. CLASSIFICAÇÃO EM TRÊS POPULAÇÕES

Outro interesse dos especialistas estava em classificar os pacientes em um dos três grupos conforme se coloca em seguida.

- POPULAÇÃO DOS PACIENTES QUE REJEITARAM O TMO EM MENOS DE 100 DIAS.
- POPULAÇÃO DOS PACIENTES QUE REJEITARAM ENTRE 100 DIAS E DOIS ANOS.
- POPULAÇÃO DOS PACIENTES QUE REJEITARAM O TMO APÓS DOIS ANOS OU QUE NÃO REJEITARAM (GRUPO CONSIDERADO NORMAL).

Com o objetivo de classificar novos pacientes em uma das três populações de tempo de rejeição, utilizou-se o banco de dados do TMO com 200 registros pré-processados e se construíram as funções de classificação e respectivas taxas de acerto, utilizando os seguintes grupos de variáveis.

- TODAS AS VARIÁVEIS PRÉ-PROCESSADAS DO PRÉ E DO PÓS-TRANSPLANTE.
- SOMENTE GRUPOS DE VARIÁVEIS DO PRÉ-TRANSPLANTE.
- SOMENTE GRUPOS DE VARIÁVEIS UTILIZADAS PELOS ESPECIALISTAS PARA FAZER ANÁLISE DE SOBREVIDA.
- SOMENTE GRUPOS DE 4 VARIÁVEIS SELECIONADAS COMO “MELHOR GRUPO PARA ESTIMAR O TEMPO ATÉ A REJEIÇÃO”.

A) UTILIZANDO TODAS AS VARIÁVEIS PROCESSADAS DO PRÉ E PÓS-TMO

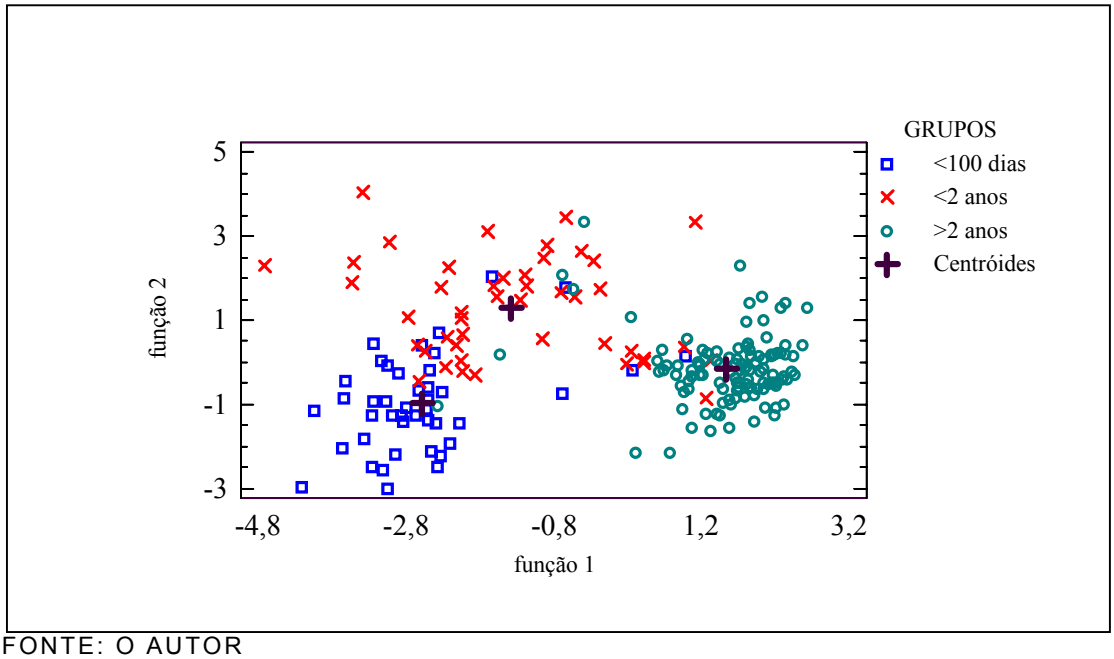
Utilizando todas as variáveis do pré e pós-transplante do Banco de Dados do TMO, modelos estatísticos padronizados de classificação foram construídos para alocar novos pacientes em uma das três classes de tempo até a rejeição do transplante. A taxa de acerto foi de 88% (ver anexo 1). Na página seguinte, o gráfico 4.5 da função de classificação, utilizando todas as variáveis, mostra que as três classes de tempo de rejeição estão bem agrupadas, facilitando a classificação de novos registros.

B) SOMENTE O GRUPO DE VARIÁVEIS DO PRÉ-TMO

Optou-se aqui por regras que se utilizam somente das variáveis coletadas dos pacientes antes do transplante (pré-tmo). Neste caso, foram construídas funções de classificação (ver anexo 1) e a taxa de acerto ficou em 63%. O gráfico 4.6, na próxima página, mostra a dispersão nas três populações.

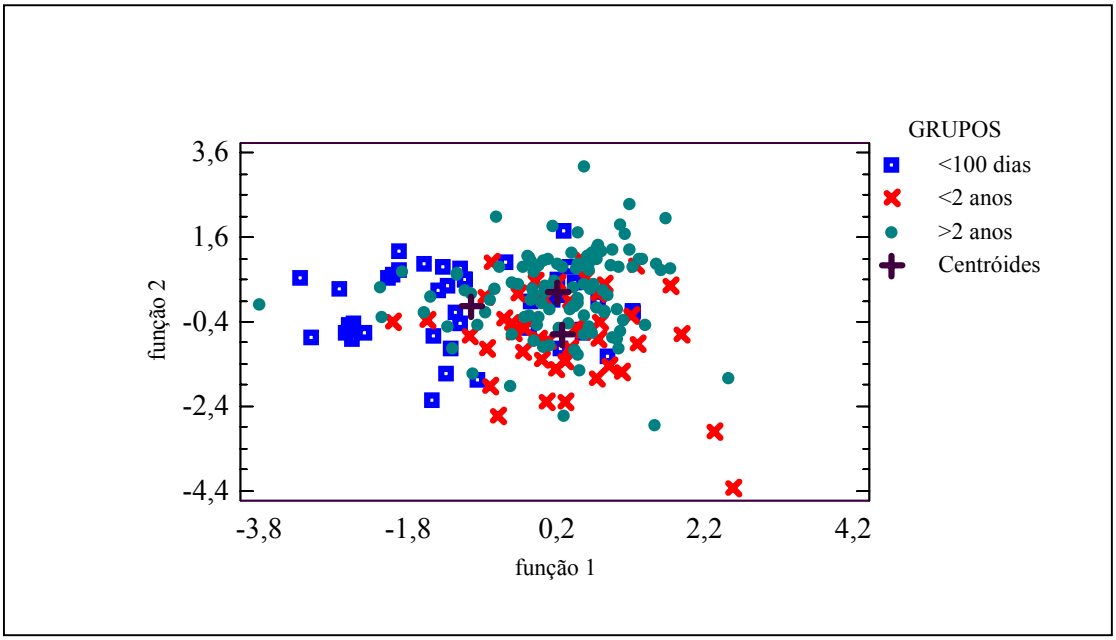
Comparando os gráficos 4.5 e 4.6 fica claro que os atributos dos pacientes no pós-transplante têm influência na rejeição.

GRÁFICO 4.5 – FUNÇÕES DISCRIMINANTES COM TODOS OS ATRIBUTOS (PRÉ E PÓS-TMO), EM TRÊS CLASSES.



FONTE: O AUTOR

GRÁFICO 4.6 – FUNÇÕES DISCRIMINANTES PARA OS ATRIBUTOS DO PRÉ-TMO, EM TRÊS CLASSES.



FONTE: O AUTOR

C) UTILIZANDO VARIÁVEIS INDICADAS PELOS ESPECIALISTAS

Segundo os especialistas em transplante, as variáveis mais significativas para o cálculo da sobrevida são o número de transfusões, número de infecções e tempo da doença (PASQUINI, 2002). Com este conhecimento, funções estatísticas de classificação foram construídas e apresentaram taxa de acerto de 57%. Portanto, aproveitando este conhecimento e buscando compor outros grupos com poucas variáveis que fornecessem uma melhor taxa de acerto, centenas de combinações foram feitas com as variáveis do pré-transplante, até se chegar a um grupo de atributos com melhor taxa de acerto, conforme item a seguir.

D) UTILIZANDO QUATRO VARIÁVEIS SELECIONADAS COMO "MELHOR GRUPO"

Na busca de melhores taxas e de modelos mais enxutos, inúmeras combinações de variáveis foram testadas. A melhor resposta foi obtida com somente quatro variáveis (karnof-pré, tipo de condicionamento pré-tmo, infecções e duração da doença), com taxa de acerto de 61%. Portanto, tomamos este conjunto de atributos para construir as funções de classificação. Assim, novos indivíduos poderão ser alocados em uma das três populações de tempo até a rejeição (ou não) com 61% de probabilidade de acerto. Os passos seguidos para a construção das funções estatísticas foram estes:

- SUMÁRIO DOS DADOS
- VETORES DAS MÉDIAS E DESVIOS PADRÕES
- MATRIZES DE COVARIÂNCIAS, CORRELAÇÕES E DE VARIAÇÕES.
- AUTOVALORES
- FUNÇÃO DE CLASSIFICAÇÃO E ESPAÇO DISCRIMINANTE
- CENTRÓIDES
- AVALIAÇÃO DO MODELO DE CLASSIFICAÇÃO
- MATRIZ DE CONFUSÃO E PROBABILIDADE A PRIORI
- ESTIMATIVAS

4.4.2.1. CLASSIFICAÇÃO DE FISHER PARA TRÊS POPULAÇÕES

Sumário estatístico das três populações.

TABELA 4.11 – SUMÁRIO ESTATÍSTICO PARA AS TRÊS POPULAÇÕES

	Médias			Medianas			Modas			Erros padrões		
VARIÁVEIS	Pop1	Pop2	Pop3	Pop1	Pop2	Pop3	Pop1	Pop2	Pop3	Pop1	Pop2	Pop3
KarnofPré	73,7	82,9	83,4	2,33	1,33	0,95	80	90	90	80	90	90
Condicionam	1,67	1,51	1,71	0,07	0,08	0,04	2	2	2	2	2	2
InfecçãoPré	1,42	1,20	1,22	0,08	0,06	0,04	1	1	1	1	1	1
DurDoença	5,91	13,4	6,96	0,67	4,20	1,01	4	4	4	2	4	2
Abo#2	1,65	1,60	1,61	0,07	0,07	0,05	2	2	2	2	2	2
AboR	2,65	2,76	2,62	0,10	0,10	0,07	3	3	3	3	3	2
CistitePós	1,19	1,11	1,04	0,06	0,05	0,02	1	1	1	1	1	1
Etiol1I2S3H	1,16	1,27	1,12	0,07	0,07	0,03	1	1	1	1	1	1
EtniaR	1,58	1,58	1,76	0,12	0,10	0,07	1	1	2	1	1	2
Evol_2viv	1,09	1,47	1,97	0,04	0,08	0,02	1	1	2	1	1	2
GrauGVHDA	0,81	0,62	0,31	0,21	0,17	0,08	0	0	0	0	0	0
GrauGVHDC	0,05	0,38	0,23	0,05	0,12	0,05	0	0	0	0	0	0
HemorrPós1	0,51	0,33	0,13	0,08	0,07	0,03	1	0	0	1	0	0
HepatPre	1,16	1,09	1,09	0,06	0,04	0,03	1	1	1	1	1	1
HipertPós	1,21	1,13	1,30	0,06	0,05	0,04	1	1	1	1	1	1
idad# (R-D)	-0,81	-2,62	-0,54	1,11	1,03	0,71	-1	-3	1	-2	-2	2
IdadeR	19,40	17,04	20,20	1,37	1,44	0,87	18	17	20	9	10	13
ImunoPre2	1,26	1,16	1,16	0,07	0,05	0,03	1	1	1	1	1	1
Imunoprof	1,95	2,00	1,99	0,03	0,00	0,01	2	2	2	2	2	2
Imunossupres2	1,05	1,40	1,04	0,03	0,07	0,02	1	1	1	1	1	1
KarnofPos	9,53	42,89	96,52	4,13	6,87	1,33	0	0	100	0	0	100
NeutInt	388	396	383	67,9	53,8	34,6	240	322	261	120	42	84
NumCelInf	3,32	3,44	3,05	0,22	0,16	0,09	3,05	3,17	2,83	3,17	2,53	2,3
PapaPos	15,19	6,62	6,63	1,84	1,04	0,42	11	5	6	15	6	6
PlaqPos	77,88	55,20	50,09	10,87	18,15	4,19	64	26	36	15	10	20
PneumPós	0,14	0,04	0,00	0,05	0,03	0,00	0	0	0	0	0	0
Sexo#2	1,51	1,47	1,42	0,08	0,08	0,05	2	1	1	2	1	1
SexoR	1,63	1,64	1,64	0,07	0,07	0,05	2	2	2	2	2	2
somaInfecPós	1,67	1,64	1,13	0,13	0,16	0,09	2	2	1	1	2	1
TransfPre	43,37	43,16	41,17	6,61	8,22	6,54	30	23	26,5	51	51	14
TratamPreDr	1,60	1,38	1,54	0,31	0,30	0,20	0	0	0	0	0	0

FONTE: O AUTOR

Os vetores médios das quatro variáveis, para cada população e global, estão na tabela 4.12 a seguir:

TABELA 4.12 - VETORES MÉDIOS DOS QUATRO ATRIBUTOS ESTRATIFICADOS EM TRÊS POPULAÇÕES, E GLOBAL

Atributos	**População 1**	**População 2**	**População 3**	**Global**
Duração da Doença	5,907	13,444	6,964	8,195
Infecções Pré	1,419	1,200	1,223	1,260
Karnof Pré (Situação %)	73,721	82,889	83,393	81,200
Tipo de Condicionamento	1,674	1,511	1,714	1,660

FONTE: O AUTOR

Os desvios padrões das variáveis para cada população e global é dado conforme tabela a seguir:

TABELA 4.13 - DESVIOS PADRÕES DOS QUATRO ATRIBUTOS ESTRATIFICADOS EM TRÊS POPULAÇÕES, E GLOBAL

Atributos	**Pop 1**	**Pop 2**	**Pop3**	**Global**
Duração da Doença	4,385	28,205	10,717	15,882
Infecções Pré	0,499	0,405	0,418	0,440
Karnof Pré (Situação %)	15,279	8,950	10,005	11,758
Tipo de Condicionamento	0,474	0,506	0,454	0,475

FONTE: O AUTOR

A seguir são mostradas as matrizes de covariâncias e de correlação.

MATRIZ DE COVARIÂNCIAS DA POPULAÇÃO UM

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	19,23	-0,603	-3,217	0,660
InfecçãoPré		0,249	-30,02	0,044
KarnofPré			233,44	-1,378
Condicionam				0,225

MATRIZ DE COVARIÂNCIAS DA POPULAÇÃO DOIS

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	795,53	-0,955	-74,27	2,563
InfecçãoPré		0,164	-0,818	0,055
KarnofPré			80,101	-1,510
Condicionam				0,256

MATRIZ DE COVARIÂNCIAS DA POPULAÇÃO TRÊS

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	116,84	-0,424	-18,43	0,313
InfecçãoPré		-0,424	0,165	-0,710
KarnofPré			106,12	-1,444
Condicionam				0,217

MATRIZ DE COVARIÂNCIA CONJUNTA

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	254,61	-0,589	-28,15	0,920
InfecçãoPré		0,1836	-1,255	0,037
KarnofPré			128,592	-1,445
Condicionam				0,228

MATRIZ DE CORRELAÇÃO CONJUNTA

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	1	-0,090	-0,1538	0,124
InfecçãoPré		1	-0,2288	0,1842
KarnofPré			1	-0,2479
Condicionam				1

A seguir são apresentadas as matrizes de variações dentro e entre os grupos.

MATRIZ DAS VARIAÇÕES DENTRO DOS GRUPOS

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	4,7612	-0,011	-0,5265	0,0172
InfecçãoPré		0,0034	-0,0235	0,0007
KarnofPré			2,4047	-0,0270
Condicionam				0,0043

MATRIZ B DAS VARIAÇÕES ENTRE OS GRUPOS

	DurDoença	InfecçãoPré	KarnofPré	Condicionam
DurDoença	35,24	-0,5815	23,6921	-0,8314
InfecçãoPré		0,0325	-1,3878	0,0102
KarnofPré			59,3066	-0,4055
Condicionam				0,0202

As funções calculadas a partir dos autovetores são usadas para discriminar novos pacientes nos três diferentes níveis de rejeição. Os coeficientes do modelo indicam quanto das variáveis independentes está sendo usada para alocar novos registros. Y2 é o modelo padronizado.

- Y1 = - 0,0258*DURDOENÇA + 0,6565*INFECÇÃOPRÉ - 0,0754*KARNOFPRÉ - 0,0708*CONDICIONAM+0,251 [4.6]

- Y2 = 0,0446 *DURDOENÇA + 0,5901*INFECÇÃOPRÉ - 0,0244*KARNOFPRÉ - 1,6314*CONDICIONAM [4.7]

Para construir o gráfico que representa o espaço discriminante nas duas dimensões, foram substituídos os dados de todos os registros nas funções discriminantes, para cada população, obtendo-se os vetores no espaço discriminante, conforme tabela 4.14. Depois, as funções Y_1 versus Y_2 foram plotadas (gráfico 4.7 na página seguinte).

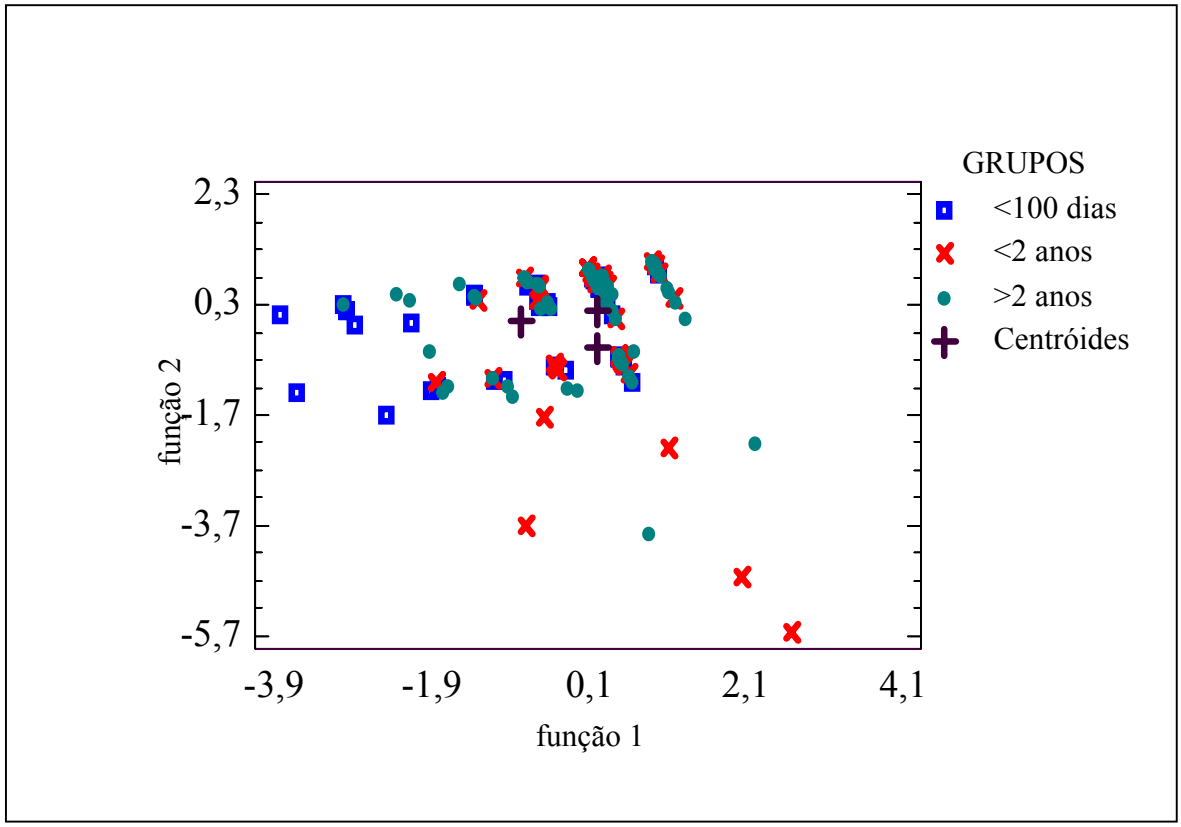
TABELA 4.14 – ESCORES FATORIAIS PARA OS QUATRO ATRIBUTOS.

Número Registro	Grupo atual	Grupo predito	Valor F1	Estimador 2	Valor F2
1	1	3*	42,01	1	41,26
2	1	3*	50,44	2	48,96
3	1	3*	50,99	2	49,67
4	1	3*	68,96	2	67,43
5	1	3*	60,96	1	60,08
6	1	1	33,59	2	32,20
7	1	3*	49,10	2	48,81
8	1	3*	48,44	2	47,97
9	1	1	22,43	3	21,58
10	1	1	5,68	3	3,16
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
191	3	3	59,10	2	57,45
192	3	3	69,07	2	67,57
193	3	3	59,32	2	57,73
194	3	3	69,40	2	67,99
195	3	3	69,20	2	67,71
196	3	3	49,99	1	48,52
197	3	3	69,73	2	68,42
198	3	3	60,21	2	58,86
199	3	3	60,65	2	59,42
200	3	3	53,55	2	52,91

FONTE: O AUTOR

NOTA: (*) Registros classificados erroneamente.

GRÁFICO 4.7 – DIAGRAMA DA FUNÇÃO DISCRIMINANTE



FONTE: O AUTOR

Centróides são as médias dos espaços discriminantes das três populações para cada uma das funções, conforme quadro a seguir:

QUADRO 4.9 – CENTRÓIDES DAS TRÊS POPULAÇÕES.

POPULAÇÕES	CENTRÓIDE (F1)	CENTRÓIDE (F2)
1	-0,7555	0,0019
2	0,2096	0,4624
3	0,2059	-0,1865

Para classificar um novo indivíduo com as regras aprendidas, substituem-se seus atributos na função discriminante, conforme equação [4.6]. Obtidos os dois valores estimados, calculam-se as distâncias d_j em relação aos centróides. Na população cuja distância d_j for menor, com taxa de acerto de 61%, alocam-se novos indivíduos, onde,

$$d_j = \sqrt{(Y_i - \mu_j)^2 + (Y_i - \mu_j)^2} \quad , \qquad \begin{cases} j = 1, 2 \text{ e } 3 \quad \text{(classes)} \\ i = 1 \text{ e } 2 \\ \mu_j \text{ são os centróides} \\ Y_i \text{ são as funções discriminantes} \end{cases}$$

Para construir a matriz de confusão, os dados dos registros atuais foram substituídos nas funções discriminantes e alocados a uma população predita. Quando as populações atuais coincidiram com as dos preditos, são contados como acerto e colocados na diagonal principal da matriz de confusão.

QUADRO 4.10 - MATRIZ DE CONFUSÃO – PARA 3 POPULAÇÕES

	GRUPO PREDITO			
GRUPO ATUAL	1	2	3	TOTAL
1	34,88%	0%	65,12%	43
2	6,67%	11,11%	82,22%	45
3	7,14%	1,79%	91,07%	112

NOTA: Porcentagem de casos classificados corretamente: 61,00%

A priori, na população atual, 21,5% dos pacientes pertencem à população um, 22,5% à população dois e 56% à população três.

QUADRO 4.11 – PROBABILIDADE A PRIORI

POPULAÇÃO	FREQÜÊNCIA RELATIVA A PRIORI
1	21,5%
2	22,5%
3	56,0%

Para o cálculo da taxa de acerto foi utilizado o método da Validação Cruzada, deixando em cada uma das dez rodadas 10% dos registros para teste e o restante para treinamento. No total, a porcentagem de registros classificados corretamente foi de 61%.

Podemos observar, na linha um do quadro 4.10 da matriz de confusão, que 34,88% dos pacientes que pertenciam ao grupo um (rejeita em menos de 100 dias) foram classificados corretamente como do grupo um. Na linha dois, 11,11% dos que pertenciam ao grupo dois (rejeita entre 100 dias e dois anos) foram classificados corretamente como do grupo dois. Na linha três da matriz, 91,07% dos pacientes que pertenciam ao grupo três (rejeita com mais de dois anos ou não rejeita) foram classificados corretamente como do grupo três.

Podemos observar na tabela 4.14 que, no primeiro registro, o paciente pertencia ao grupo um e foi classificado erroneamente como do grupo três, enquanto no registro 200, o paciente pertencia ao grupo três e foi classificado corretamente como do grupo três. No total, 61% dos registros foram classificados corretamente.

4.4.2.2. OUTROS ALGORITMOS

Na busca de melhores taxas e novas regras de aprendizagem, lançou-se mão de outros algoritmos de classificação. O pacote utilizado para executar estes algoritmos foi o *WEKA* da Universidade de Waikato

na Nova Zelândia. Os classificadores utilizados se encontram na tabela a seguir, que mostra o nome do algoritmo, a taxa de acerto e a matriz de confusão. Em seguida, alguns resultados dos algoritmos foram descritos de maneira sucinta.

TABELA 4.15 – DESEMPENHO DE ALGUNS ALGORITMOS UTILIZADOS – TRÊS POPULAÇÕES.

Algoritmos	Taxa de acerto	Matriz de Confusão Valores Preditos (%)			
		Atuais	1	2	3
J4.8 (C4.5)	50,5%	1	11,6%	2,3%	86,0%
		2	8,9%	4,4%	86,7%
		3	10,7%	5,4%	83,9%
Redes Neuronais	54,0%	1	18,6%	7%	74,4%
		2	2,2%	4,4%	93,3%
		3	9,8%	2,7%	87,5%
Decision Stump	52,2%	1	9,3%	0,0%	90,7%
		2	6,7%	0,0%	93,3%
		3	9,8%	0,0%	90,2%
Estimador Kernel	56,0%	1	18,6%	4,7%	76,7%
		2	6,7%	8,9%	84,4%
		3	8,0%	2,7%	89,3%
Logit Boost	52,5%	1	20,9%	7,0%	72,1%
		2	6,7%	6,7%	86,7%
		3	9,8%	7,1%	83,0%

FONTE: O AUTOR

DESCRIÇÃO DE ALGUNS ALGORITMOS

A) C4.5 – ÁRVORE DE PODA OU *PRUNED TREE*

O classificador WEKA C4.5 – J4.8, com parâmetros C=0,3 e M=2, mostra a árvore aprendida.

ÁRVORE APRENDIDA [4.8]

```
KarnofLanskyPre <= 60
|  DurDoenc <= 19
|  |  DurDoenc <= 3
|  |  |  KarnofLanskyPre <= 40:  => 1
|  |  |  KarnofLanskyPre > 40:   => 3
|  |  DurDoenc > 3:   => 1
|  DurDoenc > 19:     => 3
KarnofLanskyPre > 60
|  KarnofLanskyPre <= 80
|  |  DurDoenc <= 1:  => 2
|  |  DurDoenc > 1
|  |  |  InfeccaoPre = 1
|  |  |  |  Cond_cfa1 = 1
|  |  |  |  |  DurDoenc <= 8:     => 2
|  |  |  |  |  DurDoenc > 8:      => 3
|  |  |  |  Cond_cfa1 = 2:  => 3
|  |  |  InfeccaoPre = 2
|  |  |  |  KarnofLanskyPre <= 70
|  |  |  |  |  DurDoenc <= 5
|  |  |  |  |  |  DurDoenc <= 3:  => 1
|  |  |  |  |  |  DurDoenc > 3:   => 3
|  |  |  |  |  DurDoenc > 5: =>  2
|  |  |  |  KarnofLanskyPre > 70:      => 3
|  KarnofLanskyPre > 80:   => 3
```

CONCLUSÕES DO ALGORITMO C4.5

Esta árvore, de tamanho 25, gastou computacionalmente 1,41 segundo para “aprender” e testar os 200 registros, de vinte em vinte. De todos os registros, 50,50% foram classificados corretamente. O erro médio absoluto ficou em 0,38. Dos registros da classe um, 11,6% foram classificados corretamente, enquanto 4,4% dos da classe dois e

83,97% dos da classe três, conforme se vê na matriz de confusão dada na tabela 4.15.

Pode-se observar que a taxa de acerto do algoritmo C4.5 ficou bem abaixo (50,5%) da obtida pela Análise Discriminante de Fisher (61%). Outras conclusões podem ser tiradas da árvore:

- Se o paciente tiver menos de três meses da doença e *karnofPré* menor ou igual a 40, o paciente será alocado ao grupo um (grupo dos que rejeitariam em menos de 100 dias).
- Se o paciente apresentar *karnofPré* maior que 40 e tiver no máximo 19 meses da doença, o paciente será alocado ao grupo três (dos que rejeitariam em pelo menos 2 anos ou não rejeitariam).
- Se o paciente tiver pelo menos 3 meses da doença, e ainda apresentar *karnofPré* de pelo menos 60, o paciente também será alocado ao grupo três.
- Na última linha, pode-se ver que, se o paciente apresentar *karnofPré* maior que 80, independentemente da duração da doença ou de ter tido ou não infecção, o paciente será alocado ao grupo três.

B) REDES NEURONAIS

Os parâmetros utilizados aqui foram: L=0.3, M=0.2, N=1550, V=0, S=0, E=20 e NA=verdadeiro, onde L é razão de aprendizado, M são os impulsos, N é o número de interações, V é a validação, S é a semente aleatória, E é a validação cruzada, H são as camadas escondidas e NA significa se os atributos estão ou não normalizados.

CONJUNTO DE TREINAMENTO COMPLETO [4.9]

ENTRADAS	PESOS
Threshold (Nó Zero)	2,2568809
Nó 2	2,0979971
Nó 3	-3,9944326
Nó 4	-5,5617979
. . . .	. . . .
Threshold (Último nó)	-1,2786852
Duração Doença	0,6251073
Infecções Pré	-2,1377329
KarnofLansky Pré	3,8876270
Condicionamento	5,9722179

CONCLUSÕES DAS REDES NEURONAIS

A porcentagem de registros classificados corretamente foi de 54%, e o erro médio ficou em 0,3599.

Observamos também que a taxa de acerto do aprendizado em Redes Neuronais (54%) ficou abaixo da obtida pela Análise Discriminante (61%).

5. CONCLUSÃO

Juntamente com os especialistas em transplante, um grupo de atributos de pacientes doadores e receptores do TMO foi selecionado e pré-processado para ser utilizado na descoberta de conhecimento. Alguns valores faltantes foram estimados por meio de modelos de regressão múltipla. As correlações mais significativas entre as variáveis (tabela 4.2) mostraram, por exemplo, que quanto maior o número de plaquetas do paciente, maior a quantidade de papa de sangue que ele necessita (r=55,1%) no pós-transplante; ou, quanto mais *hemorragia* (r=-39,8%) e mais papa de sangue o paciente recebe depois do transplante, pior sua condição (*karnofP*ós) do paciente; ou ainda, quanto melhor está o paciente antes do transplante (*ka rnofP*ré), menos papa de sangue no pós-transplante ele recebe (r=-38%).

Com a análise fatorial, todos os atributos foram agrupados em dose fatores comuns (tabela 4.4) da seguinte maneira: nos primeiros quatro fatores ficaram agrupados os atributos do pré-transplante, no quinto atributos do pré e do pós e nos sete últimos os atributos do pós-transplante.

Quando os atributos foram classificados em grupos de rejeição, dois tipos de classificação foram utilizados: com duas populações (rejeitar ou não rejeitar o transplante) e com três populações (rejeitar em menos de cem dias, de cem dias a dois anos e em mais de dois anos). No primeiro caso, e utilizando os trinta e quatro atributos pré-processados de antes e depois do TMO, a classificação estatística de Fisher apresentou taxa de acerto de 92,5. Porém, quando utilizadas somente variáveis do pré-transplante a taxa de acerto ficou em 83%.

Com ajuda dos especialistas, várias simulações foram feitas para encontrar um número menor de atributos significativos para classificar os pacientes. Neste caso, chegou-se a um grupo com somente quatro variáveis e que apresentou taxa de acerto de 80,5%, ou seja, próxima de quando utilizadas todas as variáveis do pré-transplante. A função “aprendida” [4.2] mostra, conforme já era de conhecimento dos

especialistas, que o atributo que mais influencia na classificação dos pacientes é o condicionamento que o paciente recebe antes do transplante.

Outros algoritmos de classificação foram utilizados (tabela 4.10), como por exemplo, a Regressão Logística (RL), Redes Neuronais (RN) e J4.8, com taxas de acerto de 77%, 77% e 78,5%, respectivamente. Assim como a classificação de Fisher, a RL e as RN mostraram que o tipo de condicionamento que o paciente recebe antes do transplante tem grande influência na classificação. As regras aprendidas com o algoritmo J4.8 de Quinlan [4.5], mostram por exemplo que, se o paciente teve infecção pré-transplante e a ele for administrado o tipo de condicionamento um (cfa), então será alocado à população zero (rejeitar), ou ainda, se não teve infecção, então será alocado à população um (não rejeitar).

No segundo caso, quando a população foi dividida em três grupos, a taxa de acerto para a classificação de Fisher ficou em 61%, com maior peso para o atributo “infecções antes do transplante”. Outros algoritmos de classificação (tabela 4.15) foram utilizados, como por exemplo, RN [4.9] e J4.8 [4.8], com taxas de acerto de 54% e 50,5% respectivamente. O algoritmo J4.8 mostra na árvore aprendida por exemplo que, se o paciente tiver menos de três meses da doença e *karnofPré* menor ou igual a 40, ele será alocado ao grupo um (grupo dos que rejeitariam em menos de 100 dias); se apresentar *karnofPré* maior que 40 e tiver no máximo 19 meses da doença, será alocado ao grupo três (dos que rejeitariam em pelo menos 2 anos ou não rejeitariam); se o paciente tiver pelo menos 3 meses da doença, e *karnofPré* em pelo menos 60, também será alocado ao grupo três.

As regras “aprendidas” foram repassadas aos especialistas do setor de transplante e ao grupo de estatísticos da IBMTR (*International Bone Marrow Transplantation Registry* - serviço norte-americano que concentra dados estatísticos de todos os serviços de transplante do mundo) para análise e implantação.

6. RECOMENDAÇÕES

Dado a situação do paciente num determinado momento depende do seu estado imediatamente anterior, recomenda-se a aplicação das ferramentas do Processo Estocástico. Recomenda-se também que seja ampliada a base de dados, utilizando registros de pacientes de outros centros de transplantes.

ANEXOS

ANEXO 1 – ALGUNS RESULTADOS DAS FUNÇÕES DE CLASSIFICAÇÃO

FUNÇÃO DISCRIMINANTES PARA DUAS POPULAÇÕES, UTILIZANDO AS 17 VARIÁVEIS DO PRÉ-TMO.

Y= -0,14*KarnofLanskyPre –0,28*DurDoenc +0,18*NeutInt +0,22*InfecPre +0,11*TransfPre –0,025*Etiologia1I2S3H - 0,18*ImunoPre –0,1*Sexo#Dif2 +0,68*CondCfa +0,05*SexoR +0,45*IdadeTMOrec +0,003*EtniaRec -0,14*AboR – 0,435*idR_idD –0,05*Abo#2 +0,10*HepatitePre +0,26*TratamPreDra - 0,0218.

FUNÇÃO DISCRIMINANTE PARA TRÊS POPULAÇÕES, UTILIZANDO TODAS AS VARIÁVEIS DO PRÉ E PÓS-TMO.

Y1= -0,0324447*SexoR + 0,115924*Idade - 0,0325692*EtniaRec +0,0886424*AboR + 0,0882339*id# - 0,0104788*Abo#2 + 0,11324*KarnofPre +0,172075*DurDoenca + 0,0463341*HepatitePre - 0,105689*NeutInt + 0,393191*Cond_cfa1 + 0,0142146*InfeccaoPre - 0,134295*ImunoPre_2 +0,112218*TratamPre_Dra - 0,146798*TransfPrevia + 0,0310656*Sexo#2 - 0,117305*Etiologia1I2S3H + 0,025097*Imunoprofil - 0,0147803*NumCelInf - 0,0466577*CistitePos - 0,225105*HemorPos_1 + 0,0660365*GrauGVHDA + 0,186187*GrauGVHDC + 0,0586777*HipertPos - 0,185257*PapaPos + 0,0602214*PlaqPos + 1,04885*KarnofPos - 0,425393*Imunoss2 + 0,0132.

MATRIZ DE CONFUSÃO - VARIÁVEIS DO PRÉ+PÓS E DO PRÉ-TMO.

	GRUPO PREDITO					
GRUPO ATUAL	1 pré+pós	1 pré	2 pré+pós	2 pré	3 pré+pós	3 pré
1	86,05%	39,53%	6,98%	2,33%	6,98%	58,14%
2	13,33%	6,67%	71,11%	26,67%	15,56%	66,67%
3	0,89%	7,14%	3,57%	6,25%	95,54%	86,61%

ANEXO 2 – O TRANSPLANTE DE MEDULA ÓSSEA - TMO

Este anexo foi dividido em onze partes para apresentar o que é o TMO, suas etapas, os tipos e funções, as complicações, a doença enxerto contra hospedeiro e venoclusiva, as infecções, o crescimento e desenvolvimento, a recuperação do sistema imunológico e as funções do TMO.

HISTÓRICO DO TMO NO PARANÁ

O Serviço de Transplante de Medula Óssea (TMO) do Hospital de Clínicas da Universidade Federal do Paraná iniciou suas atividades em 1979 sendo pioneiro na América Latina. Está hoje entre os 16 serviços do mundo que realizam mais de 100 procedimentos por ano. Em abril de 1998 foi realizado o milésimo procedimento do Serviço e, atualmente, responde por mais de 40% dos transplantes alogênicos realizados no país. Foi pioneiro também na realização de transplantes, usando células tronco hematopoéticas obtidas de cordão umbilical e de medula óssea, obtidas de doador não-aparentado.

É afiliado ao *International Bone Marrow Transplantation Registry* - IBMTR, serviço norte-americano que concentra dados estatísticos de todos os serviços de transplante do mundo, além de manter contato freqüente com os grandes centros de transplante tais como St. Jude em Memphis e Fred Hutchinson em Seattle nos Estados Unidos.

Atualmente o Serviço de Transplante de Medula Óssea ocupa o 15º andar do Hospital de Clínicas, uma unidade especialmente construída para este fim. Conta com um sistema de ar filtrado sob pressão, o que mantém o ar isento de partículas. Ocupa também o 4º andar do prédio de ambulatório que foi devidamente reformado para o atendimento de pacientes que não mais necessitam de regime de internação.

São 140 profissionais entre médicos, enfermeiros, assistentes sociais, fisioterapeutas, dentistas, psicólogos, nutricionistas e administradores, compondo uma equipe multidisciplinar do maior nível técnico e totalmente atualizada nas modernas técnicas de transplante.

QUE É TMO?

Denomina-se transplante de medula óssea o procedimento terapêutico em que se realiza a infusão venosa de células do tecido hematopoético, com a finalidade de restabelecimento da hematopoese após aplasia medular, seja ela de causa benigna primária (ex. anemia aplásica), causada por neoplasia maligna (ex. leucemias e linfomas) ou ainda relacionada ao tratamento realizado para estas neoplasias (exemplo: radioterapia).

Dessa maneira, o papel do transplante de medula óssea varia de acordo com a sua indicação, desde o restabelecimento da hematopoese na anemia aplásica ao suporte hematopoético para viabilizar a administração de regimes de altas doses de quimioterapia para o tratamento das neoplasias malignas.

As células progenitoras do sistema hematopoético que farão o re-povoamento medular podem ser obtidas basicamente de três fontes: células diretamente aspiradas da medula óssea, célula tronco periféricas (“stem cell”) mobilizadas do compartimento medular para o sangue periférico, ou células progenitoras do cordão umbilical.

HISTÓRICO DO TMO

O primeiro relato de administração de células hematopoiéticas com finalidade terapêutica, data de 1891, quando Brown-Sequard e D'Arsonaval administraram medula óssea por via oral em pacientes com anemia causada por leucemia. Em 1937, Schretzenmayr, foi o primeiro a administrar por via intramuscular, medula óssea fresca autóloga ou alogênica em pacientes com anemias relacionadas à malária ou

infestação por helmintos. Em 1940, Marrison e Samwick descreveram pacientes com anemia aplásica que se recuperaram após três infusões intramedulares de apenas 13 ml de aspirado de medula óssea dos seus irmãos. Experimentalmente, Jacobson e colaboradores demonstraram que era possível evitar aplasia medular em camundongos que recebiam radioterapia, com a infusão de células esplênicas. Em trabalhos subseqüentes, Lorenz, Congdon e Uphoff (1952) além de Lorenz e Congdon (1954), relataram a eficácia terapêutica da suspensão de células de medula óssea no tratamento de anemia aplásica.

As décadas de 50 e 60 foram marcadas por frustrações e desapontamentos, a maioria dos transplantes era em doentes terminais que não tinham sobrevida suficiente para avaliação da eficácia do enxerto. Os enxertos em que havia sucesso na pega, geralmente resultavam em reação enxerto-hospedeiro ou septicemias, sempre letais.

Em 1957, Goren descobriu aloantígenos relacionados ao complexo de histocompatibilidade em camundongos (denominado H2) e Dausset, em 1964 descreveu o antígeno leucocitário humano (HLA-A2 human leukocyte antigen). Somente em 1972, Thomas e colaboradores, relataram o primeiro transplante de medula óssea alogênico com sucesso para anemia aplásica com doador HLA (Human Leucocyte Antigen).genotipicamente idêntico.

TIPOS DE TRANSPLANTES DE MEDULA ÓSSEA

Os transplantes de medula óssea podem ser divididos basicamente em dois tipos: o alogênico e o transplante autólogo onde não existe doador e as células utilizadas são provenientes do próprio paciente. Quando realizado entre irmãos gêmeos, o transplante é denominado de singênico. Existem ainda os transplantes alogênicos entre pessoas não aparentadas (de outras famílias), onde a célula doada pode vir de um "banco" de medula óssea. Mais recentemente descobriu-se que o sangue existente no cordão umbilical é muito rico

em células progenitoras da medula óssea (células denominadas de CD34+), abrindo uma nova possibilidade de transplantes, denominado de transplante com células de cordão.

Para a realização do transplante alogênico, é fundamental que o doador apresente o HLA bastante parecido ao do receptor, de acordo com uma análise dos loci A, B, DR e DQ, que constituem o que se denominam de tipagem de HLA classe 1 e 2.

Infelizmente, grande número de pacientes não dispõe de doadores compatíveis e este problema estimulou o desenvolvimento do transplante não aparentado. O maior obstáculo deste método é a contaminação da medula óssea por células neoplásicas; que inviabilizaria o tratamento. Na tentativa de solucionar este problema, foram desenvolvidos vários métodos de purificação destas células, que são conhecidos como “purning” da medula óssea. Os métodos mais utilizados empregam a quimioterapia in vitro e anticorpos monoclonais.

ETAPAS DO TRANSPLANTE

O procedimento pode ser dividido nas seguintes fases:

- MOBILIZAÇÃO E COLETA DA MEDULA ÓSSEA OU CÉLULAS TRONCO.
- CONDICIONAMENTO COM QUIMIOTERAPIA COM OU SEM RADIOTERAPIA.
- PEGA E RECUPERAÇÃO MEDULAR.

COLETA DA MEDULA ÓSSEA OU CÉLULA TRONDO

A coleta das células pode ser realizada de duas formas: cirúrgica, por meio de múltiplas punções aspirativas de medula (preferencialmente em crista ilíaca posterior) em ambiente cirúrgico, sob anestesia geral ou peridural e a coleta por meio do sangue periférico (aférese de células mononucleares). Por meio da abordagem cirúrgica, o objetivo é coletar entre 10 a 15 ml de medula óssea por Kg de peso do doador ou receptor, ou o equivalente a 3×10^8 de células medulares nucleadas por Kg de peso. O material coletado é

armazenado em recipientes contendo anticoagulante (heparina) e conservantes específicos. Por meio deste método, a medula pode ficar congelada por cerca de 10 anos.

Há cerca de 8 anos, descobriu-se que as células da medula óssea poderiam ser mobilizadas para o sangue periférico por meio da estimulação com fatores de crescimento hematopoéticos. O mais importante e mais amplamente utilizado é o fator de crescimento de colônia de granulócitos, que estimula as células tronco a "saírem" do compartimento medular e circularem no sangue periférico, onde são coletadas por meio de aférese. Dependendo da indicação, a medula óssea ou as células-tronco podem ser "purificadas" para remoção das células indesejadas (neoplásicas ou linfócitos T). O material coletado é criopreservado em nitrogênio líquido e, no momento oportuno, infundido por meio de veia central.

CONDICIONAMENTO

A quimioterapia de altas doses associada ou não à radioterapia, administradas previamente à infusão da medula óssea, têm três objetivos:

a) Erradicação da medula doente do receptor.
b) Erradicação do sistema imune do receptor para que as células do doador sejam aceitas.
c) Proporcionar "espaço" para a nova medula.

Existem inúmeros regimes de condicionamento com radioterapia de corpo inteiro associados ou não à quimioterapia. As drogas mais usadas são: ciclofosfamida, BCNU, cisplatina, carboplatina, etoposide, thiotepa, bussulfano, melfalan e ifosfamida. A toxicidade relacionada ao condicionamento varia de acordo com a combinação de drogas utilizada. A quase totalidade dos pacientes desenvolve mucosite em algum momento pós-condicionamento, associada ou não a febre.

RECUPERAÇÃO MEDULAR

Após o regime de condicionamento, o paciente passa por um período de aplasia medular em que é necessário suporte hemoterápico adequado.

Estratégias para evitar aloimunização (sensibilização do HLA) incluem a utilização de um doador único (para plaquetas e glóbulos), irradiação dos hemoderivados e utilização de filtro de leucócitos. Estes procedimentos permitem o aumento da meia-vida das plaquetas infundidas e otimiza sua eficácia, relacionando-se também com a diminuição do risco de soro-conversão e infecção por CMV. É recomendado que pacientes com sorologia negativa para CMV, que receberam medula também negativa, recebam somente produtos de doadores soro-negativos.

O GVHD pode ser efetivamente prevenido pela irradiação de produtos sangüíneos antes da transfusão. Estudos recentes sugerem que a dose de 1500 a 2000 cGy pode reduzir "mitogen-responsive lymphocytes" por 5 a 6 logs comparados com produtos não irradiados. Os grupos e subgrupos sangüíneos, tanto do doador como do receptor, bem como títulos de anticorpos devem ser pesquisados. Diante da incompatibilidade ABO, pode ser feita plasmaférese do receptor ou remoção das hemácias da medula a ser infundida.

Do ponto de vista prático, é necessário que se mantenha um nível de hemoglobina acima de 10.0 g/dl e 20.000 plaquetas/mm3. Contagens abaixo desses níveis indicam a necessidade de transfusão. Denomina-se "pega" medular, o momento onde a contagem plaquetária é mantida acima de 20.000/mm^3 por três dias seguidos sem a necessidade de transfusão e os granulócitos estão acima de 500/mm^3, também por 3 dias consecutivos.

COMPLICAÇÕES DO TMO

As principais complicações do transplante de medula óssea podem ser divididas de uma maneira didática em:

a) Associadas ao regime de condicionamento
b) Associadas à infusão das células tronco ou medula óssea
c) Doença do enxerto contra hospedeiro
d) Doença venoclusiva
e) Sangramentos e
f) Infecções.

TOXICIDADES RELACIONADAS AO REGIME DE CONDICIONAMENTO

Toxicidade cardíaca: cerca de 90% dos regimes contendo ciclofosfamida apresentam um quadro de pequenas alterações eletrocardiográficas, arritmias supraventriculares ou pericardites sem comprometimento hemodinâmico. Porém, 5 a 10% dos regimes contendo ciclofosfamida apresentam eletrocardiograma com baixa voltagem, insuficiência cardíaca progressiva e até pericardite com ou sem tamponamento. Carmustine (BCNU) é outro agente quimioterápico muito usado em regimes de condicionamento e tem sido associado à toxicidade cardíaca aguda.

Aparelho urinário: a toxicidade urotelial é uma das complicações mais freqüentes da ciclofosfamida em altas doses. A acroleína, um dos metabólitos finais da ciclofosfamida, quando exposta ao urotélio, resulta em hiperemia e até ulceração da mucosa com hemorragia e necrose focal. Estratégias para prevenção de cistite hemorrágica, consistem na hiper-hidratação e administração. O tratamento de cistite hemorrágica severa requer correção de plaquetopenia, hidratação generosa e irrigação da bexiga.

Toxicidade renal: insuficiência renal depois do transplante de medula pode ser resultado de nefrotoxicidade direta da radioterapia ou dos agentes quimioterápicos. Estes incluem cisplatina, ifosfamida e ciclofosfamida. Além disso, lise tumoral, depleção do volume intravascular e outras drogas nefrotóxicas tais como anfotericina B,

aminoglicosídeos e ciclosporina, também podem causar insuficiência renal. Usualmente, a insuficiência renal após o transplante de medula óssea é o resultado de múltiplos insultos ao rim. Em um estudo de 272 pacientes (Fred Hutchinson Cancer Research Center), 53% tinham dobrado o nível basal de creatinina e 24% necessitaram de hemodiálise.

Toxicidade pulmonar: pneumonia não infecciosa ou relacionada ao regime ocorre em cerca de 8 a 18% dos pacientes que recebem transplante de medula óssea. Esta incidência parece não diferir entre as modalidades de transplante alogênico, autólogo ou singênico. O quadro clínico clássico consiste em dispnéia, infiltrada pulmonar difusa, tosse seca e hipoxemia. Trata-se da causa mais comum de infiltrado pulmonar difuso nas primeiras quatro semanas após o transplante e é mais comum em transplantes para malignidades hematológicas. Radioterapia, e uma variedade de agentes quimioterápicos, tais como a ciclofosfamida, bussulfano e BCNU são diretamente tóxicos para os pulmões. O lavado broncoalveolar é o procedimento diagnóstico inicial para diferenciar pneumonias não infecciosas de pneumonias por citomegalovírus. Quando este procedimento não for esclarecedor, há indicação de biópsia pulmonar. O tratamento consiste em suporte ventilatório e administração de altas doses de corticóides, porém quando há necessidade de ventilação mecânica, o prognóstico é muito pobre.

Mucosite. Vários estudos têm demonstrado que a incidência de mucosite excede 90% dos casos de transplante de medula. Durante o regime de condicionamento ocorre xerostomia e após a infusão da medula, a mucosa começa progressivamente a ulcerar. A dor é em geral muito importante, com necessidade de administração de analgésicos opióides. A resolução deste quadro ocorre quando há recuperação medular. Regimes contendo irradiação de corpo inteiro, bussulfano, etoposide e thiotepa são mais freqüentemente associados com mucosite. A superinfecção da mucosa oral com fungos, bactérias, ou vírus são comuns e pode influenciar na severidade e na duração da

mucosite. Espécies de cândida e vírus do herpes simples são os patógenos mais comumente isolados em pacientes com mucosite prolongada.

Pele. Irradiação de corpo inteiro e a maioria dos quimioterápicos podem causar toxicidade cutânea. Eritema generalizado e hiperpigmentação da pele são comuns em pacientes que recebem altas doses de radioterapia. Drogas citotóxicas com significante toxicidade cutânea são citosina arabinosídeo, thiotepa, BCNU, bussulfano e etoposide. A Biópsia de pele demonstra uma variedade de alterações inflamatórias. Em casos severos, o uso de corticóide sistêmico pode ser indicado para obter controle antiinflamatório.

Irradiação de corpo inteiro. Os efeitos colaterais agudos mais importantes são náuseas e vômitos. Com menor freqüência pode-se observar síncope, edema das glândulas salivares e fadiga.

COMPLICAÇÕES RELACIONADAS À INFUSÃO DE CÉLULAS TRONCO OU MEDULA ÓSSEA:

São pouco comuns, podendo ocorrer micro êmbolos pulmonares, reações alérgicas, hemólise por incompatibilidade ABO e sobrecarga de volume.

DOENÇA ENXERTO CONTRA HOSPEDEIRO (DECH):

A identidade imunológica de um indivíduo é expressa por proteínas da superfície celular codificadas pelo sistema de histocompatibilidade que, nos humanos, é denominado de H.L.A. Por meio destas proteínas, o sistema imune reconhece tecidos invasores e os destrói, mecanismo pelo qual ocorre a reação do hospedeiro contra o enxerto em tecidos transplantados. No transplante de medula óssea alogênico, ocorre o inverso, o tecido transplantado em questão, imunológicamente competente, pode reconhecer o hospedeiro como "proteínas invasoras" e iniciar a reação enxerto contra hospedeiro.

Em 1966, Billingham postulou que, para haver doença enxerto contra hospedeiro, três requisitos devem ser preenchidos:

1. O enxerto deve conter células imunologicamente competentes (linfócitos T).
2. O receptor deve expressar antígenos teciduais que não estão presentes no doador do transplante.
3. O receptor deve ser incapaz de realizar uma resposta imune para destruir as células transplantadas.

A doença enxerto contra hospedeiro pode ser observada em formas clínico-patológicas: aguda e crônica.

DECH AGUDA

A maioria dos transplantes de medula óssea alogênicos, sem profilaxia imunossupressora, desenvolverá DECH. DECH aguda pode ocorrer entre os primeiros dias até dois meses após o transplante. A incidência varia de 10 a 80% dependendo do grau de histocompatibilidade, número de células T no enxerto, idade do paciente e regime profilático. Os órgãos mais acometidos são pele, intestino e fígado. DECH aguda ocorre primeiro e mais comumente na pele, caracterizando-se por:

Rash maculopapular pruriginoso inicialmente nas palmas das mãos, plantas dos pés e orelhas,

Freqüentemente progride como eritrodermia em todo corpo com formação de bolhas e descamação em casos severos.

Manifestações hepáticas e gastrointestinais geralmente aparecem mais tarde e raramente representam o primeiro sinal de DECH. Os sintomas intestinais constituem, inicialmente, anorexia, náuseas e vômitos, que podem progredir para diarréia, dor abdominal e até íleo paralítico. DECH hepática é caracterizada por hiperbilirrubinemia, aumento da fosfatase alcalina e aminotransferases, alterações da coagulação e, em casos mais severos, falência hepática.

Estágio Clínico de DECH

	Pele	Fígado	Intestino
+	Eritema máculopapular <25% SAC*	Bilirrubina 2-3mg/dl	Diarréia 500-1000ml/dia
++	Eritema máculopapular 25-50% SAC*	Bilirrubina 3-6mg/dl	Diarréia 1500-1000ml/dia
+++	Eritrodermia Generalizada	Bilirrubina 6-15mg/dl	Diarréia >1500ml/dia
++++	Descamação e bolha	Bilirrubina >15mg/dl	Dor ou íleo

Nota: *SAC: superfície de área corpórea.

- A epiderme e seus folículos são danificados e até destruídos.
- Ductos biliares menores são profundamente afetados com ruptura segmentar.
- A destruição das criptas intestinais resulta em ulcerações mucosas que podem ser localizadas ou difusas. A gravidade depende da graduação da DECH, graus I e II apresentam baixa morbidade, e graus III e IV, a mortalidade é alta.

DECH CRÔNICA

Foi inicialmente definida como síndrome da DECH presente 100 dias depois do transplante de medula óssea, porém pode ser observada após 40 a 50 dias do transplante. Sua incidência varia de 30 a 60%. DECH crônica pode ser limitada ou extensa, de acordo com os seguintes critérios:

Tipo de Doença	Extensão da Doença
Limitada	Envolvimento localizado da pele, disfunção hepática, ou ambos
Extensa	Envolvimento generalizado da pele, Envolvimento localizado da pele ou disfunção hepática associada a um dos seguintes:

	• Hepatite crônica agressiva, necrose em ponte ou cirrose, • Acometimento ocular, • Envolvimento de glândulas salivares, • Envolvimento das mucosas (biópsia de lábio) • Envolvimento de outros órgãos alvo.

Os órgãos mais acometidos são: pele (80%), fígado (50%), olhos (30%), intestinos (30%) e boca (80%). DECH crônica em pele pode apresentar-se como líquen plano, placas, dermatites papuloescamosas, descamações, despigmentações e vitiligo. Destruição dos anexos podem levar à alopécia e à onicodisplasia. As formas mais severas podem assemelhar-se à esclerodermia. DECH crônica hepática freqüentemente lembra a aguda e raramente evolui para cirrose. Mucosite severa na cavidade oral e esôfago podem resultar em perda de peso e desnutrição. O envolvimento gastrointestinal é freqüente. DECH crônica pode provocar destruição linfocítica das glândulas exócrinas, causando atrofia e secura das superfícies mucosas, geralmente acometendo olhos, boca, vias aéreas, pele e esôfago. O sistema hematopoético também pode ser atingido e trombocitopenia é fator prognóstico desfavorável. Características patológicas:

1. Sistema imune: involução do epitélio tímico, depleção de linfócitos e ausência de centros germinativos secundários em linfonodos.

2. Pele: atrofia da epiderme (alterações características de líquen plano), esclerose da derme e fibrose da epiderme.

3. Gastrointestinal: processos inflamatórios localizados nas mucosas e formações severas em esôfago e intestino delgado.

4. Fígado: semelhante à DECH aguda, porém mais intensa, com alterações crônicas tais como obliteração dos ductos biliares e colestase hepatocelular.

5. Bronquiolite obliterante: semelhante à rejeição do transplante pulmonar, são geralmente consideradas como manifestação de DECH crônica embora sua patogênese seja controversa.

O seguimento da DECH crônica pode ser determinado por prognósticos desfavoráveis: ataque progressivo, alterações liquenóides da pele, níveis elevados de bilirrubinas, trombocitopenia persistente e falência de resposta à terapia por 9 meses. Espera-se que 70% dos pacientes com nenhum destes fatores sobrevivam, comparados com 20% de sobrevida nos pacientes que apresentam 2 ou mais fatores de risco.

DECH EM TRANSPLANTE SINGÊNICO

São geralmente autolimitadas, afetam predominantemente a pele. Embora o grau de severidade possa ser II ou III, geralmente é rapidamente resolvido com administração de glicocorticóides sem risco de vida.

PROFILAXIA E TRATAMENTO DA DECH

Para melhor entendimento da terapêutica adotada, é importante lembrar que basicamente a DECH é dividida em duas fases:

- Fase aferente: os tecidos do hospedeiro ativam os linfócitos T do doador e as citoquinas envolvidas são a interleucina 1 e 2.
- Fase eferente: proliferação clonal dos linfócitos T, recrutamento de células adicionais e ataque às células alvo.

Baseado nestes princípios imunofisiopatológicos, a profilaxia tem como objetivo combater a fase aferente. Existem duas estratégias.

1. Bloqueio da ativação dos linfócitos T por meio de glicocorticóides, ciclosporina e methotrexate sendo que as duas últimas drogas compõem a associação profilática mais difundida.
2. Remoção dos linfócitos T da medula óssea a ser infundida; por meio de método físico ou por meio de anticorpos monoclonais contra células T.

Esta estratégia resulta em redução substancial na incidência e severidade da DECH. Infelizmente, o uso de medula óssea com

depleção de linfócitos T é associada com as taxas mais altas de falha de pega do enxerto, e com a incidência aumentada de recidiva de alguns tipos de leucemia, particularmente LMC; este fato parece estar relacionado ao efeito enxerto contra leucemia.

O tratamento da DECH aguda consiste basicamente no uso de corticoesteróides e suporte clínico. Outras modalidades terapêuticas como globulina antitimócitos, ciclosporina, anticorpos monoclonais (anti CD3 e anti interleucina 2), tem sido utilizados somente para doenças esteróides resistentes. O tratamento de suporte consiste em prevenir ou limitar a exposição a organismos infecciosos. Lesões cutâneas "abertas" devem ser abordadas como queimaduras graves e conjuntivites severas podem necessitar de tratamento tópico. No envolvimento gastrointestinal, o uso de antibióticos não absorvidos, nutrição parenteral e reposição hídrica são úteis. Os pacientes podem beneficiar-se de antibioticoterapia profilática e antifúngica.

O tratamento da DECH crônica consiste no uso de corticosteróides, sendo a prednisona a droga de eleição. Talidomida, azatioprina, clofazimina, micofenolato mofetil e PUVA (8-metoxipsoralen associado a irradiação ultravioleta A) são drogas que parecem efetivas tanto na profilaxia com no tratamento da DECH.

DOENÇA VENOCLUSIVA (VOD)

A doença venoclusiva é uma síndrome clínica caracterizada por icterícia, hepatomegalia e retenção de líquidos (ganho de peso). É patologia relacionada à toxicidade hepática pós-condicionamento e geralmente ocorre nas primeiras semanas após o transplante. Esta síndrome é decorrente do dano das células endoteliais, sinusóides e hepatócitos ao redor das vênulas hepáticos terminais. Pode variar em severidade, de leve e reversível à fatal, associada com falência de múltiplos órgãos. O diagnóstico é geralmente clínico pela tríade icterícia, hepatomegalia e ganho de peso e ocorre cerca de 8 a 10 dias após o final do condicionamento. Ocasionalmente, dor no hipocôndrio

direito ocorre devido à distensão da cápsula hepática e simultaneamente, observa-se retenção de sódio com resultante ganho de peso que está relacionado ao desenvolvimento de hipertensão intrasinuisoidal devido à obstrução do fluxo sangüíneo hepático. A hiperbilirrubinemia é mais tardia, o edema periférico ocorre em 60% e ascite em 20% dos casos.

Devem ser feitos diagnósticos diferenciais com infiltração fúngica, DECH hiperaguda, injúrias hepáticas causadas por outras medicações e colangite lenta. A incidência de VOD severa é maior em regimes de condicionamento, contendo irradiação de corpo inteiro e ciclofosfamida (CFA) e BCV (BCNU e etoposide). Regimes com bussulfano também têm sido associados com alta incidência de VOD. Hepatite pré-transplante, febre, resposta inflamatória durante a administração da quimioterapia e infusão de medula de doador com HLA não totalmente compatível, são os fatores de risco de VOD mais importantes.

O tratamento é primariamente de suporte, visando evitar balanço hídrico positivo Se necessário devem ser utilizados diuréticos. Dada a evidência de que fatores de coagulação são depositados no espaço subendotelial das vênulas danificadas, trombólise tem sido proposta e estudos não controlados têm demonstrado que o ativador do plasminogênio tecidual recombinante (RTPA) e a heparina podem ser efetivos. A infusão de prostaglandina E_1, que tem efeito vasodilatador e prostaglandina antitrombótico, parece ser também efetiva no tratamento da DVO.

COMPLICAÇÕES TARDIAS DO TMO

A partir dos anos 80, a sobrevida dos pacientes que receberam transplante de medula óssea vem aumentando.

Conseqüentemente, a importância dos efeitos tardios relacionados ao transplante de medula óssea vem ganhando destaque, principalmente em pacientes pediátricos. Os regimes de condicionamento com irradiação de corpo inteiro e quimioterapia com

agentes alquilantes são associados com um risco aumentado de complicações malignas e não malignas.

CRESCIMENTO E DESENVOLVIMENTO DO TMO

Retardamento do crescimento é um problema comum em crianças irradiadas. A produção de GH é reduzida em 90% das crianças onde a radioterapia craniana foi incluída, comparada com 40% daquelas que não receberam radioterapia craniana pré-transplante. Puberdade é retardada ou ausente em crianças irradiadas. Somente uma minoria das meninas atinge a menarca espontaneamente, a maioria necessita de reposição de hormônios femininos. Em contraste, os meninos geralmente recuperam a função das células de *Leydig* com produção de testosterona. O desenvolvimento das características sexuais secundárias é atrasado pela presença de DECH crônica.

FUNÇÕES DO TMO

As funções do TMO são divididas em Gonodal e Reprodutiva Pós-Puberal, Pulmonar e Músculo Esquelética.

GONADAL E REPRODUTIVA PÓS-PUBERAL

A função gonadal em pacientes receptores de transplante é deteriorada pelo efeito direto da quimioterapia e radiação nas gônadas. Todas as mulheres irradiadas desenvolvem falência ovariana primária e menos que 10% apresentam recuperação entre 3 e 7 anos.

Espermatogênese é persistentemente ausente na maioria dos homens irradiados, mas a fertilidade pode ocorrer vários anos depois. Anormalidades tireoideanas são observadas em cerca de 40% dos pacientes transplantados, tanto hipo como hipertireoidismo.

FUNÇÃO PULMONAR

Disfunções respiratórias não são raras em pacientes transplantados. As patologias mais comuns são pneumonite intersticial de início tardio e bronquiolite obliterante. As alterações restritivas mais severas ocorrem em pacientes com pneumonite intersticial prévia. Drogas, radioterapia e DECH crônica podem contribuir na patogênese.

FUNÇÃO MÚSCULO ESQUELÉTICO

Miosites, monoartrites ou poliartrites podem ocorrer em pacientes com DECH crônica. Distrofia ou atrofia muscular pode ser resultado de DECH crônica ou uso de corticóides por tempo prolongado.

Doença óssea: cerca de 10% dos pacientes transplantados evoluem com osteonecrose asséptica. A osteoporose é comumente detectada seguida de fraturas patológicas; pode também estar relacionada à menopausa precoce pós-transplante, DECH crônica e uso de corticóide por tempo prolongado.

Função neurológica: polineuropatia é ocasionalmente vista nos pacientes transplantados. As disfunções dos nervos periféricos são geralmente associadas à infecção por herpes zoster. Leucoencefalopatia multifocal tem sido observada em crianças.

Anormalidades oftalmológicas: cerca de 80% dos pacientes que receberam irradiação de corpo inteiro em dose única, desenvolvem catarata em 6 anos; entretanto somente 20% dos pacientes que receberam irradiação de corpo inteiro em doses fracionadas ou dose única baixa. Ceratoconjuntivite crônica, candidíase, CMV e outros patógenos podem provocar coriorretinite.

Aparelho urinário: deterioração da função glomerular persistente é observada em alguns pacientes. Insuficiência renal de início tardio com anemia, hipertensão e retenção de fluídos pode ocorrer. Cistite hemorrágica tardia e câncer de bexiga têm sido descritos.

Neoplasias malignas secundárias: doença maligna primária, quimioterapia, radioterapia e imunossupressão são condições que podem aumentar o risco de neoplasias secundárias. Em vários estudos, foi observado que as incidências de linfomas não Hodgkin, leucemias, glioblastoma multiforme e carcinoma hepatocelular foram significativamente aumentadas.

RECUPERAÇÃO DO SISTEMA IMUNOLÓGICO E HEMATOPOÉTICO APÓS TRANSPLANTE DE MEDULA ÓSSEA

Com o regime de condicionamento, o paciente perde seu sistema linfo-hematopoético. A infusão de medula óssea fornece um novo sistema imune.

O tempo de recuperação é variável, os leucócitos geralmente reaparecem em 2 a 3 semanas. O número de neutrófilos aumenta mais rapidamente que os linfócitos. Reticulócitos seguem o mesmo padrão dos leucócitos. A recuperação plaquetária é a mais lenta. A medula óssea geralmente é hipocelular no 2º a 3º mês pós-transplante. Normalização é observada em torno do 3º a 6º mês pós-transplante.

A função oxidativa dos neutrófilos é normalizada rapidamente, porém a quimiotaxia permanece reduzida por vários meses. Fatores que influenciam para neutropenia após a enxertia são ocorrência de DECH, infecções (particularmente viral), uso de drogas mielossupressoras, incluindo methotrexate, cotrimoxazol e ganciclovir.

A incompatibilidade ABO tem pouca ou nenhuma influência na taxa de recuperação leucocitária e plaquetária; porém a reconstituição eritrocitária fica atrasada. O uso de células tronco-periféricas proporciona uma enxertia mais rápida que quando utilizamos células tronco provenientes da medula óssea. A falha de enxerto precoce ocorre em 1% dos transplantes de medula óssea para leucemias condicionados com irradiação de corpo inteiro e ciclofosfamida. A incidência de falência de enxerto parece estar relacionada com o grau de disparidade do HLA.

Fatores estimulantes de colônias de granulócitos tem mostrado acelerar a recuperação dos granulócitos, assim como fator estimulante de colônias de granulócitos e monócitos, proporcionando uma redução no tempo de internação, incidência de febre e uso de antibióticos.

A contagem de linfócitos totais retorna ao normal ao redor de 12 semanas. Porém linfócitos T4 permanecem em taxas reduzidas entre 6 e 12 semanas, enquanto os linfócitos T8 retornam rapidamente aos valores normais e, freqüentemente, permanecem elevados por longo tempo. Há assim uma inversão característica na relação CD4:CD8 vista em TMO alogênicos, autólogos e singênicos.

O número total de células B retorna ao normal após um mês de TMO. As concentrações séricas de imunoglobulinas IgG e IgM retornam ao normal aproximadamente 9 meses pós-transplante.

DECH aguda ou crônica e uso de drogas imunossupressoras são fatores que lentificam a recuperação imunológica.

Devido à recuperação lenta da imunidade após TMO, vacinação convencional com vírus vivo ou atenuado não é recomendada. Entretanto tem sido descritas a eficácia e segurança do uso de vacinas com vírus atenuados para sarampo, parotidite e rubéola.

INFECÇÕES NO TMO

As infecções são as maiores causas de morbidade e mortalidade subseqüentes à quimioterapia de alta dose com ou sem transplante de medula óssea (TMO).

A medula óssea é o órgão mais lesado com quimioterapia intensiva e a granulocitopenia ocorre secundariamente a este dano, o que predispõe estes pacientes ao desenvolvimento de infecções severas. É observado que o nível de granulocitopenia determina o risco dessas infecções. O dano a outros órgãos, causada pela doença ou pelo tratamento com terapia intensiva em conjunto com granulocitopenia resulta em sítios, onde bactérias e fungos podem não só causar infecções localizadas, mas também servir de entrada na

circulação sangüínea, levando à bacteremias e fungemias. Com a ausência de granulócitos e sem antibioticoterapia apropriada, a bacteremia e fungemia levam a choque séptico e morte.

FATORES DE RISCO

As áreas mais propensas à infecção em granulocitopênicos são:

- Trato gastrointestinal, incluindo cavidade oral.
- Trato respiratório.
- Pele.

Os candidatos a este procedimento devem ser submetidos a exame odontológico antes de iniciar tratamento, pois a doença periodontal, pode ser um sitio de infecção quando a granulocitopenia se desenvolve.

A maioria dos compostos usados causa toxicidade mucosa e a perda da integridade dessa barreira, resulta em sítio de infecção; a mucosite pode ser potenciada pela reativação do vírus herpes simples.

A mucosa do trato gastrointestinal também é atingida e este dano leva a sintomas como dor torácica com esofagite, diarréia, cólicas abdominais com lesão no esôfago, intestino delgado e grosso. O tempo é coincidente entre o dano da medula óssea e mucosa. A área perianal também é sitio freqüente de infecção.

Geralmente os pacientes que recebem quimioterapia em alta dose com ou sem TMO, possuem cateter de longa duração como Hickman, Broviac ou Groschong, que facilitam a administração de quimioterapia, antibióticos, hemoderivados, fluídos e outras terapias de suporte. O número de venopunções e o risco de infecção de pele, porém, cria um sítio potencial de infecção.

Sinais de hiperemia e dor no sitio da saída ou túnel do cateter pode ser fonte de infecção, associada com bacteremias.

O trato respiratório pode ser outro sitio de infecção e a sinusite não é incomum nestes pacientes. Dessa forma, deve-se fazer

rotineiramente tomografia de seios da face antes do TMO, para evitar posterior avaliação de sitio de infecção.

Os pulmões são também atingidos, pois alterações na produção de muco e na função ciliar, aumentam o risco de infecções. Dessa forma os pacientes que recebem quimioterapia intensiva são de risco para infecção não só pela supressão da medula óssea e granulocitopenia, mas por alterações nos mecanismos de imunidade celular e humoral.

ANEXO 3 – ENDEREÇOS ELETRÔNICOS PESQUISADOS

AnswerTree SPSS - http://www.spss.com
ChartWorks Server Visual Mining - http://www.visualmining.com
Clementine SPSS - http://www.spss.com
Data Mining - www.dct.ufms.br/ mzanusso/dm.html.
DataBase Mining Marksman HNC Software Inc. - http://www.hncs.com
DataMiner 3D Dimension 5 - http://www.dimension5.sk
DataMite Logic Programming Associates - http://www.lpa.co.uk
DBMiner DBMiner Technology - http://www.dbminer.com
Enterprise Miner SAS – http://www.sas.com
FuzzyDecisionDesk Fuzzy Logik Systeme GmbH - http://www.fuzzy.de
Hospital Virtual Brasileiro – www.hospvirt.org.br/transmedula
Intelligent Miner IBM – http://www1.ibmlink.ibm.com
Knowledge Access Suite Information Discovery - www.datamining.com
Minitab www.cs.waikato.ac.nz/ml/weka
NeuroGenetic Optimizer BioComp Systems
Neuronal Connection SPSS Inc. - http://www.spss.com
Neuronal Network Browser Apreal, Inc. - http://members.aol.com/apreal
Neuronal Techno Neuronal Technologies - http://www.neuronalt.com
NeuronalNet Tutor www.attg.com
Neuronalyst For Excel www.cheshirreng.com
NeuroShell Ward Systems Group, Inc - http://www.wardsystems.com
NeuroSolutions NeuroDimension, Inc. - http://www.nd.com
Nuggets *Data Mining* Technologies, Inc. - http://www.data-mine.com
Parallel Visual Explorer IBM - http://www.ibm.com
Pattern Recognition Workbench Unica - http://www.unica-usa.com
Redes Neuronais - http://www.cpdee.ufmg.br/~apbraga/nns/nnmain.htm
SAS System SAS Institute - http://www.sas.com
Simulator Matrica - http://www.matrica.org -
S-Plus MathSoft - http://www.mathsoft.com
Statgrhaphics www.statgraphics.com
TMO -HC-UFPR www.ufpr.br/hosp/tmo
VisualMine Artificial Intelligence Software SpA - www.visualmine.com

REFERÊNCIAS BIBLIOGRÁFICAS

ARNS, Maria Terezinha Steiner. **Uma Metodologia para o reconhecimento de padrões multivariados com resposta dicotômica.** UFSC, Florianópolis, 1995.

ASSIS, Francisco de. **Discussão em Portal Brasileiro**. 1999**.**

BASTOS, Rogério C. **Avaliação de desempenhos educacionais**: uma Abordagem usando Conjuntos difusos. Tese de doutorado. Universidade Federal de Santa Catarina. Departamento de engenharia de produção, 1994.

BRAGA, A. P.; CARVALHO, A. C. P & LUDERMIR, T. B. **Fundamentos de redes neurais artificiais**: XI escola brasileira de computação, 1998.

BISCHOP, Christopher M. **Neural networks for pattern recognition**. Oxford university press, 1995.

CARVALHO, L. A. V. **Data mining**: A mineração de dados no marketing, medicina, engenharia e administração. São Paulo. Érica, 2001.

CHAVES Neto, Anselmo. **Apostila de análise multivariada**. Pós-graduação em métodos numéricos em engenharia – UFPR. Curitiba, 2002.

COSTA, E.; BENTO, C. **Extração de conhecimento em banco de dados**. DEI-FCTUC, 2001.

DINIZ, C. A. R.; LOUZADA-NETO, F**. Data mining**: Uma introdução. Departamento de estatística – UFSCAR, São Carlos, 2000.

DOBSON, A. **An introduction to statistical modeling.** Chapman and Hill, London, 1983.

DUDA, R. A.; HART, P. E. **Pattern classification and scene analysis.** New Work: Wiley & Sons, 1973.

FAYYAD, U. et al. **From data mining to knowledge discovery:** an overview**,** AAAI Press, 1996.

FAUSETT, Laurence. **Fundamentals of neural networks:** Institute of technology. Ed. Prentice Hall. Flórida 2000.

FELDENS, A. M. **Descoberta de conhecimento em bases de dados e mineração de dados.** Instituto de informática, UFRGS, RS, 2002.

FREITAS, Alex. Pontifícia Universidade Católica. Disponível em: <www.ppgia.pucpr.br> Acesso em 25 Nov. 2001.

FRIEDMAN, J. **Multivariate adaptive regression splines** (with discussion). Annals of statistics, 1991.

GEMAN, S.; BINENSTOCK, E. and DOURSAT, R. **Neural networks and the bias/variance dilemma**. Neural computation. Press, 1992)

GOEBEL, Michael and GRUENWALD, Le**. A survey of data mining and knowledge discovery software tools.** ACM SIGKDD, v.1, 1999.

GOLDBERG, David E. **Genetic algorithms in search optimization and machine learning**. Alabama: Addinson-Wesley, 1989.

HASSE, Mozart. **Mineração de dados usando algoritmos genéticos**. Curitiba: dissertação apresentada no curso de mestrado em informática – UFPR. Curitiba, 2000.

HASTIE, T. & TIBSHIRANI, R. **Discriminant analysis by gaussian mixtures.** Journal of the Royal Statistics Society, Séries B, 1994.

HASTIE, T.; BUJA, A. and TIBSHIRANI, R. **Flexible discriminant analysis by optimal scoring.** Journal of the American Statistics Association, 1994.

HECHT & NIELSEN. **Theory of the back propagation neural network.** In IEEE Int. Joint Conf. On Neural Network, 1989.

HOLSHEIMER, M.; KERSTEN, M. and SIEBES, A. **Data survey: searching the nuggets in parallel in:** FAYYAD, U. M. et al. (Eds.) Advances in knowledge discovery and data mining, AAAI Press, 1996.

JOHNSON, Richard A. & WICHERN, Dean W. **Applied multivariate statistical analysis**. Fourth edition. New Jersey: Prentice Hall, 1998.

KENNEDY, R.L. **Solving data mining problems with pattern recognation.** Prentice Hall, 1999.

KFOURI, Eduardo. **Somente a inteligência pode combater as fraudes,** 2003.

KLIR, G. J. & YUAN, B. **Fuzzy sets and fuzzy logic - theory and applications**. Prentice Hall, 1995.

KLEIN, J.P, at al. **Statistical methods for the analysis and presentation of the results of bone marrow transplantation -** IBMTR, USA, 2001.

LACHENBRUCH, P. A & MICKEY, M. R. **Estimation of error in discriminate analysis.** Thecnometrics, 1968.

LEÃO, Beatriz de Faria: **Inteligência artificial aplicada à medicina**. Instituto cardiológico do Rio Grande do Sul, 2000.

LOYOLA, M. **Data mining**. Endereço do DCT da UFMS. Disponível em: <http://www.dct.ufms.br> Acesso em: 10 Set. 2001.

MICHALEWICZ, Zbigniew. **Genetic algorithms + data structures = evolution programs.** Third revised and extended edition. S.l.: Springer, s.d., 2000.

NAVEGA, S. **Princípios essenciais da *data mining*.** SP, 2002.

NEVES, Heliz Regina Alves. **Implementação de um sistema especialista para determinar elegibilidade e prioridade em transplante de medula óssea**. Curitiba: Dissertação de mestrado apresentada no curso de mestrado em informática aplicada – PUC, 2000.

NIEVOLA, Júlio. Pontifícia Universidade Católica. Disponível em: <www.ppgia.pucpr.br> Acesso em 25 Fev. Curitiba, 2002.

NORTON, M. Jay. **Knowledge discovery in databases.** Library Trends, 1999.

PASQUINI, R.: Universidade Federal do Paraná. **Que é tmo?.** Serviço de transplante de medula óssea – TMO. Disponível em: <www.ufpr.br/tmo> Acesso em 05 Jan. 2001.

QUINLAN, J. Ross. **C4.5**: **Programs for machine learning**. San Mateo, California: Morgan Kaufmann Publishers, 1993.

RAITTZ, Roberto Tadeu. **Free associative neurons – FAN**. Uma Abordagem para reconhecimento de Padrões. Disponível em: <http://www.eps.ufsc.br/disserta98/raittz> Dissertação obtida pela Universidade Federal de Santa Catarina, 1997

ROSS, Timoty J. **Fuzzy logic with engeneering applications**. McGrow-Hill inc., 1995.

RUMELHART, D. E. **Neuroconcience and connectionist theory.** Edition Hardcover, 1988.

RUMELHART, D. E. **Learning internal representation by error propagation.** Edition Hardcover, 1986.

SCHALKOFF, R. **Pattern regognition, statistical, structural and neural approaches.** John Wiley, 1992.

Universidade Federal de Minas Gerais. Departamento de engenharia elétrica. **Redes neurais artificiais**. Endereço do grupo de pesquisas em RNA. Disponível em: <www.aesetorial.com.br/tecnologia/artigos> Acesso em 25 Jan. 2003.

WITTEN, I. H & FRANK, E. **Data mining**: **Practical machine learning tools and techniques with java implementations**. Morgan Kaufmann. California, 1999.

ZANUSSO, M. **Data mining**. DCT - UFMS. Disponível em: <http://www.dct.ufms.br> Acesso em: 15 nov. 2002.