Resumo

Compreender as razões que levam uma mulher com mais de 60 anos, residente em Curitiba, estar com hipertensão é uma tarefa difícil e delicada. Esta investigação visa identificar prováveis fatores que influenciem a aparição da doença. Estudos apontam a relação cintura-quadril, índice de massa corporal, bem como hábitos de alimentação e prática de exercícios físicos como sendo influenciadores desse mal. Outro fator está relacionado à mulher não saber que está com hipertensão e demorando assim para iniciar o tratamento. Sabendo isso, a pesquisadora Maressa Priscila Krause fez um levantamento de dados junto às unidades de saúde de Curitiba, solicitando a esse laboratório assessoria na análise dos mesmos. A metodologia utilizada consiste na coleta de dados, através de questionário específico, contemplando características individuais das mulheres. Esses dados foram submetidos a uma análise estatística, subsidiando a comparação de métodos estatísticos como análise de cluster, árvores de decisão, regressão logística, análise linear discriminante e redes neurais artificiais. Os referidos métodos procuram estimar os fatores que têm impacto sobre a probabilidade de se ter hipertensão. Através desses métodos confirmou-se que a hipertensão está associada a obesidade, aumento da idade e falta de exercícios físicos.

Lista de Tabelas

Lista de Figuras

Sumário

1 INTRODUÇÃO

O risco que a elevação da pressão arterial representa para o sistema cardiovascular e outros órgãos é bem conhecido ([30]).

A hipertensão arterial constitui um dos problemas de saúde de maior prevalência na atualidade. Conforme [31], estima-se que a hipertensão arterial atinja aproximadamente 22% da população brasileira acima de vinte anos, sendo responsável por 80% dos casos de acidente cérebro vascular, 60% dos casos de infarto agudo do miocárdio e 40% das aposentadorias precoces, além de significar um custo de 475 milhões de reais gastos com 1,1 milhão de internações por ano.

Em 1998, o Ministério da Saúde/Brasil já indicava a existência de 8.100.000 hipertensos, com níveis tencionais iguais ou superiores 160/95mmHg (milímetros de mercúrio), número que dobraria se o critério de diagnóstico de hipertensão arterial sistólica fosse reduzido para níveis de PA na faixa de 140/90mmHg. Dados em [26] motram que mais da metade desses hipertensos não sabia do diagnóstico e, portanto, não eram atendidos pelo sistema de saúde para tratamento da doença e mais da metade daqueles que conheciam o diagnóstico abandonavam o tratamento por diversos motivos. Estes dados mostram uma baixa adesão da população hipertensa ao tratamento, bem como sua cobertura pelo sistema público de saúde.

Segundo [25], a hipertensão arterial é citada como uma síndrome clínica caracterizada pela elevação da pressão arterial a níveis iguais ou superiores a 140mmHg de pressão sistólica e/ou 90mmHg de diastólica. Geralmente, é uma doença silenciosa: não dói, não provoca sintomas, entretanto, pode matar. Quando ocorrem sintomas, já decorrem de complicações.

Em face a todos os problemas que a hipertensão representa, [20] ressalta a importância do tratamento antihipertensivo na redução da morbidade e mortalidade cardiovasculares, principalmente na prevenção de acidentes vasculares, insuficiência cardíaca e renal. O controle da hipertensão arterial inicia-se com a detecção e observação contínua, não devendo ser diagnosticada com base em uma única medida da pressão arterial.

Dados citados em [16] mostram que a hipertensão arterial tem maior freqüência de diagnóstico quanto maior a idade do examinando

Com todos estes elementos, este estudo tem por objetivo identificar os principais fatores de risco relacionados a hipertensão em mulheres com mais de 60 anos residentes em Curitiba, utilizando métodos estatísticos multivariados. Primeiramente, iremos descrever um pouco sobre a doença que nos motivou a realizar esse estudo.

1.1 TEMA DO ESTUDO

A hipertensão arterial é o aumento desproporcionado dos níveis da pressão em relação, principalmente, à idade. A pressão arterial normal num adulto alcança um valor máximo de 140mmHg (milímetros de mercúrio) e mínimo de 90mmHg segundo as informações em [9].

A Pressão sangüínea é a força gerada pela contração do coração para manter adequada e constante a circulação do sangue através dos vasos. Para superar a resistência oferecida por quilômetros de estreitos vasos sangüíneos e para que o sangue chegue aos tecidos com pressão residual suficiente para a troca de substâncias químicas, o coração deve manter um nível mínimo de pressão dentro do sistema circulatório (ver [23]).

Hipertensão arterial é a pressão arterial acima de 140x90mmHg (milímetros de mercúrio) em adultos com mais de 18 anos, medida em repouso de quinze minutos e confirmada em três vezes consecutivas conforme descrito em [5]. A Tabela 1, retirada de [9] contém as variações de pressão sistólica e diastólica e possíveis ações a serem tomadas.

Nível	Pressão arterial sistólica	Pressão arterial diastólica	Ação a tomar
Hipotensão	inferior a 100	inferior a 60	check-up médico
Valores normais	entre 100 e 140	entre 60 e 90	auto-mediç ao
Hipertensão limite	entre 140 e 160	entre 90 e 100	check-up médico
Hipertensão moderada	entre 160 e 180	entre 100 e 110	consultar o médico
Hipertensão grave	superior a 180	superior a 110	consultar o médico com urgência
Hipertensão sistólica específica	superior a 140	inferior a 90	consultar o médico

Tabela 1: Variações da Pressão

2 REVISÃO DE LITERATURA

2.1 ANÁLISE MULTIVARIADA

Investigação científica é um processo interativo do conhecimento. Objetivos relacionados de um fenômeno social ou físico devem ser descritos e então verificados por análise de dados. Por sua vez, uma análise de dados coletados por experimentação ou observação podem normalmente sugerir uma explicação modificada do fenômeno. Durante este processo interativo do conhecimento, variáveis são freqüentemente adicionadas ou removidas do estudo. Assim, as complexidades da maioria dos fenômenos requerem uma investigação para coletar observações de diferentes variáveis. Estamos preocupados com métodos estatísticos definidos para obter informações desse gênero de grupos de dados. Estes dados incluem simultaneamente medidas ligadas a muitas variáveis, esta metodologia é chamada Análise Multivariada.[13]

Em [13], os objetivos de investigações científicas, para os quais métodos múltivariados são usados, incluem os seguintes:

1. Redução de dados ou Simplificação estrutural: Os fenômenos estudados são representados tão simples como possíveis sem sacrificar informações valiosas. É esperado que permita fazer interpretações facilmente.

2. Classificação e agrupamento: Grupos de objetos parecidos ou variáveis são criados, baseados por características calculadas. Alternativamente, regras para classificação de objetos bem definidas.

3. Investigação de dependência entre variáveis: A natureza da relação entre as variáveis é de interesse. Todas as variáveis são mutuamente independentes ou existe uma ou mais dependente das outras? Se sim, como?

4. Predição: Relacionamento entre variáveis deve ser determinado com a finalidade de predizer os valores de uma ou o mais variáveis com base na observação das outras variáveis.

5. Construção e teste de hipóteses: Hipóteses estatísticas específicas, formuladas em relação aos parâmetros de populações multivariadas são testadas. Isso pode ser feito para validar suposições ou reforçar convicções prévias.

2.1.1 ANÁLISE DE CLUSTER

Análise de Cluster é um conjunto de técnicas utilizadas na identificação de padrões de comportamento em bancos de dados através da formação de grupos homogêneos (ver [8]).

O objetivo da analise de cluster é agrupar as observações semelhantes de forma que cada grupo seja homogêneo internamente e sejam diferentes entre si.

[24] enumera os passos para a análise de cluster da seguinte forma :

1. Definir medidas de similaridade entre os objetos.
2. Decidir qual a técnica de clusterização será utilizada: hierárquica ou não hierárquica.
3. Decidir sobre o método de clusterização para a técnica que já foi selecionada (exemplo: método do centróide para a técnica hierárquica de clusterização).
4. Decidir sobre o número de clusters a serem feitos.
5. Interpretar a solução final.

MEDIDAS DE SIMILARIDADE E DISSIMILARIDADE

Cada objeto é representado por um ponto no espaço n-dimensional e, portanto, pode ser agrupado com outros que estejam próximos e mais se assemelham a ele. Em algumas ocasiões o interesse está no agrupamento de variáveis e consideramos estas no espaço p-dimensional.

Há dois tipos de medidas para identificar pontos semelhantes: medidas de similaridade (quanto maior o valor, maior a semelhança entre os objetos) e medidas de dissimilaridade (quanto maior o valor, mais diferentes são os objetos).[8]

Freqüentemente há muitas subjetividades na escolha das medidas de similaridade. [13] coloca importantes considerações para a escolha, incluindo a natureza das variáveis (discreta, contínua, binária) ou escalas de medida (nominal, ordinal, intervalar, intervalar, razão) e conhecimento do assunto.

Quando objetos (unidades ou itens) são agrupados, a proximidade é normalmente indicada em alguns tipos de distância. Por outro lado, variáveis são usualmente agrupados com base nos coeficientes de correlação ou como medidas de associação.[13]

Na seqüência apresentamos as distâncias em função do tipo de variável.

1. Variáveis Quantitativas

   As distâncias são as medidas de dissimilaridade mais utilizadas no estudo de bancos de dados com variáveis quantitativas.[8]

   Uma medida d_{ik} representa uma distância entre os pontos i e k se:

   - $d_{ik} \geq 0$ para qualquer escolha de i e k;
   - $d_{ii} = 0$;

- $d_{ik} = d_{ki}$;
- $d_{ik} \leq d_{im} + d_{mk}$;

Algumas distancias utilizadas são:

- Distância Euclidiana
  A idéia básica é considerar cada observação como sendo um ponto num espaço euclidiano e, desse modo, a formula nos dá a distância física entre os pontos.[8]

$$d_{ik} = d_{ik}^{(2)} = \sqrt{(x_i - x_k)^T (x_i - x_k)} = \sqrt{\sum_{j=1}^{p} (X_{ij} - X_{kj})^2} \tag{1}$$

- Distância de Manhattan *(city block)*

$$d_{ik}^{(1)} = \sum_{j=1}^{p} |X_{ij} - X_{kj}| \tag{2}$$

2. Variáveis Categorizadas

   O tratamento básico das variáveis qualitativas consiste na codificação de

   suas respostas através de variáveis indicadoras *(dummies)*.

   Para construção das medidas de similaridade ou dissimilaridade, resume-se as informações de dois indivíduos conforme indicado na tabela 2.

	individuo k		
individuo i	1	0	Total
1	a	b	a+b
0	c	d	c+d
Total	a+c	d	b+d

Tabela 2: Construção das medidas de similaridade e dissimilaridade

Dois indivíduos com comportamento semelhante terão valores elevados na diagonal principal. Por outro lado, se o comportamento for diferente, os valores mais elevados estarão na diagonal secundaria. Baseado nesse raciocınio sugere-se, respectivamente, as seguintes medidas de similaridade e dissimilaridade:

$$s_{ik} sik = \frac{a+d}{m}$$

e

$$\delta_{ik} = \frac{b+c}{m}$$

Note que s_{ik} é a proporção de concordâncias entre as variáveis indicadoras e δ_{ik} a de discordâncias.(conforme [8])

Métodos Hierárquicos Aglomerativos

Nesses métodos os agrupamentos são formados a partir de uma matriz de similaridade ou dissimilaridade. Num primeiro passo a matriz é utilizada para identificar o par de objetos que mais se parece. A partir desse instante esse par é agrupado e será considerado como sendo um único objeto. Isso requer que se defina uma nova matriz; em seguida identifica-se o par mais semelhante, que formará um novo grupo, e assim sucessivamente até que todos os objetos estejam reunidos num mesmo grupo. Através da análise do histórico do agrupamento, pode-se definir a posteriori o número de grupos existentes nos dados. Maiores detalhes sobre estes métodos podem ser vistos em [8].

O que diferencia esses métodos é a regra para a redefinição da matriz de similaridade (ou dissimilaridade) a cada união de pares de objetos. Abaixo, são exemplificados alguns deles.

- Método do vizinho mais próximo: A distância considerada é a menor distância entre um elemento de G1 (grupo 1) e um elemento de G2 (grupo 2), ou seja,

$$d(G1, G2) =_{i \in G1} min_{k \in G2} d_{ik} \tag{3}$$

- Método do vizinho mais longe: Define-se a distância como a maior distância entre um elemento de G1 e um elemento de G2, ou seja,

$$d(G1, G2) =_{i \in G1} max_{k \in G2} d_{ik} \tag{4}$$

- Método das médias das distâncias: Nesse método calcula-se a média das distâncias entre os elementos de G1 e os de G2.

$$d(G1, G2) = \sum_{i \in G1} \sum_{k \in G2} \frac{d_{ik}}{g1g2} \tag{5}$$

- Método da centróide: Esse método define a coordenada de cada grupo como sendo a média das coordenadas de seus objetos. Uma vez obtida essa coordenada, denominada centróide, a distância entre os grupos é obtida através do cálculo das distâncias entre as centróides.

- Método de Ward: A alocação de um elemento a um grupo é feita de modo a minimizar uma medida de homogeneidade interna.

  A cada etapa do método de Ward, busca-se unir objetos que tornem os agrupamentos formados os mais homogêneos possível. A medida de homogeneidade utilizada baseia-se na partição da soma de quadrados total de uma análise de variância.

Comparação dos métodos hierárquicos

O método do vizinho mais longe tende a formar grupos mais homogêneos do que o método do vizinho mais próximo. Isso se deve ao fato de ser um critério bastante rigoroso, pois a distância pequena entre dois grupos implica na proximidade de todos os elementos desse grupo.

O método de Ward, é atraente por basear-se numa medida com forte apalo estatístico e por gerar grupos que, assim como os do método do vizinho mais longe, possuem alta homogeneidade interna.

Métodos Não Hierárquicos ou de Partição

Tecnicas não-hieráquicas são utilizadas para formar k *clusters*[1] itens ou objetos.

- Método das k-médias

  Este método exige que se estipule a priori o número de grupos que devem ser gerados.

  Em uma versão simples, o processo é composto por estes três passos:

  1. Partição dos itens em k grupos iniciais;
  2. Prosseguir com a lista de itens, colocando cada item no grupo cuja média (centróide) está mais próximo.(Usualmente calcula-se a distância Euclidiana com observações padronizadas ou não.) O centróide é recalculado para o grupo que recebeu um novo item e para o grupo que perdeu o item.
  3. Repete-se o segundo passo até que não restem relocações a serem feitas.

- Método das k-medóides

  O método das k-medóides é um método de partição baseado numa matrz de ditâncias entre objetos. A medóide de um grupo é definida como o membro do grupo que possui

[1] A definição do número ideal de clusters é um dos problemas destes métodos.

a menor distância euclidiana média em relação aos demais membros do grupo. O critério de qualidade utilizado no método consiste na minimização da soma das distâncias entre as observações e as respectivas medóides.

Sendo k o número de grupos a serem formados, o algoritmo busca identificar k vetores observados que sejam representativos dos grupos (medóides). Desse modo, o critério de qualidade do método é dado por:

$$C = \sum_{j=1}^{n} C_j C_j = min_{1 \leq i \leq k} d[m_i, j] \tag{6}$$

em que C é o critério de qualidade e $d[m_i, j]$ representa a distância entre a medóide $i(m_i)$ e a observação j. Uma vez identificados esses pontos, aloca-se cada objeto ao grupo de medóide mais próxima.

O PROBLEMA DA CLASSIFICAÇÃO

Quando se utiliza um determinado método para classificar objetos em grupos, podem ser gerados erros de determinada grandeza, ou seja, um elemento x pode ser classificado em uma determinada população, mas na realidade ele não pertence a essa população na qual foi designada. Esse tipo de erro pode ocorrer principalmente na região fronteira entre as populações, cujas características dos elementos (observações) sejam muito semelhantes.

Um outro tipo de problema de classificação é o custo do erro de classificação. Quando há duas populações π_1 e π_2, o erro de classificar um objeto de π_1 como pertencente a uma classe π_2, talvez seja maior do que o de classificar um objeto de π_2 na classe π_1.

TSUCHIYA(2002) formula matematicamente o problema da classificação da seguinte maneira: sejam $f_1(x)$ e $f_2(x)$ funções densidades de probabilidade associadas ao vetor aleatório X de dimensão $p \times 1$, respectivamente das populações π_1 e π_2. Um objeto, com medida x associada, deve ser alocado para π_1 ou π_2. Seja Ω o espaço amostral, isto é, o conjunto de possíveis observações de x. Seja R_1 o conjunto de valores de x que são classificados como objetos de π_1 e $R_2 = \Omega - R_1$ os restantes valores de x que são classificados como objetos de π_2 e, consequentemente, $\Omega = R_1 \bigcup R_2$.

Desta forma, todo o objeto deve ser associado a uma e somente uma das duas populações, os conjuntos R_1 e R_2 são mutuamente excludentes.

Considere $p(2|1)$,a probabilidade de classificar, incorretamente, um objeto como de π_2 quando na verdade ele pertence à população π_1. E $p(1|2)$ a probabilidade de cometer o outro possível erro. As duas expressões para estas probabilidades são obtidas como :

$$p(2|1) = p(x \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(x)dx \tag{7}$$

$$p(1|2) = p(x \in R_1|\pi_2) = \int_{R_1=\Omega-R_2} f_2(x)dx \tag{8}$$

Seja P_1 a probabilidade a priori de π_1 e P_2 a probabilidade a priori de π_2 , então $P1 + P2 = 1$.

A probabilidade de classificar corretamente, ou incorretamente os objetos pode ser derivada como o produto entre as probabilidades a priori e probabilidades condicionais. Desta maneira, há 4 possíveis desfechos

P(classificada corretamente como π_1) = P(observação vem de π_1 e é classificada corretamente como π_1)

$$p(x \in R_1|\pi_1) * P(\pi_1) = p(1|1) * P_1 \tag{9}$$

P(classificada incorretamente como π_1) = P(observação vem de π_2 e é classificada incorretamente como π_1) =

$$p(x \in R_1|\pi_2) * P(\pi_2) = p(1|2) * P_2 \tag{10}$$

P(classificada corretamente como π_2) = P(observação vem de π_2 e é classificada corretamente como π_2) =

$$p(x \in R_2|\pi_2) * P(\pi_2) = p(2|2) * P_2 \tag{11}$$

P(classificada incorretamente como π_2) = P(observação vem de π_1 e é classificada corretamente como π_2) =

$$p(x \in R_2|\pi_1)P(\pi_1) = p(2|1) * P_1 \tag{12}$$

Um bom modelo de classificação deve minimizar as probabilidades de classificações incorretas.

2.1.2 ÁRVORES DE DECISÃO

As árvores de decisão constituem métodos para problemas de regressão e classificação muito utilizados recentemente. A grande vantagem presente nestes procedimentos é a facilidade na interpretação do modelo final.

Uma árvore de decisão consiste de uma hierarquia de nós internos e externos que são conectados por ramos. O nó interno, também conhecido como nó decisório ou nó intermediário, é a unidade de tomada de decisão que avalia através de teste lógico qual será o próximo nó descendente ou filho. Em contraste, um nó externo (não tem

nó descendente), também conhecido como folha ou nó terminal, está associado a um rótulo ou a um valor.

Em geral, a construção de uma árvore de decisão é feita pelo seguinte procedimento: apresenta-se um conjunto de dados ao nó inicial (ou nó raiz que também é um nó interno) da árvore; dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos e este procedimento é repetido até que um nó terminal seja alcançado. A repetição deste procedimento caracteriza a recursividade da árvore de decisão.

No caso das árvores de decisão binária, cada nó intermediário divide-se exatamente em dois nós descendentes: o nó esquerdo e o nó direito. Quando os dados satisfazem o teste lógico do nó intermediário seguem para o nó esquerdo e quando não satisfazem seguem para o nó direito. Logo, uma decisão é sempre interpretada como verdadeira ou falsa.

O teste lógico é aplicado a uma das covariáveis presentes no estudo. A escolha da covariável é feita em função do seu poder de discriminação entre as classes da variável resposta.

O aprendizado de uma árvore de decisão é supervisionado, ou seja, o método aproxima funções alvo de valor discreto, na qual a função aprendida é representada por uma árvore de decisão. As árvores treinadas podem ser representadas como um conjunto de regras "Se-Então"para melhoria da compreensão e interpretação.

Árvores de decisão usadas para problemas de classificação são chamadas de Árvores de Classificação. Nas árvores de classificação cada nó terminal ou folha contém um rótulo que indica a classe predita para um determinado conjunto de dados. Nesse tipo de árvore pode existir dois ou mais nós terminais com a mesma classe.

Árvores de decisão usadas para problemas de regressão são chamadas de Árvores de Regressão. Nas árvores de regressão, cada nó terminal ou folha contém uma constante (geralmente, uma média) ou uma equação para o valor previsto de um determinado conjunto de dados.

Embora muitos algoritmos para construção de árvores de decisão tenham surgido na área de Aprendizado de Máquina (*Machine Learning*), a monografia CART escrita por BREIMAN et al. (1984) apresenta este método sobre a ótica da Estatística.

2.1.3 REGRESSÃO LOGíSTICA

Como o objetivo deste trabalho é identificar os fatores de risco para a presença de hipertensão, que é uma a variável dicotômica (0 = ausência e 1 = presença), escolhemos o método de regressão logística para modelar este fenômeno. Abaixo descreveremos partes principais da teoria que sustenta nossa escolha.

De acordo com [19], a regressão logística tem se constituído num dos principais métodos de modelagem estatística de dados. Mesmo quando a resposta de interesse não é

originalmente do tipo binário, alguns pesquisadores têm dicotomizado a resposta de modo que a probabilidade de sucesso possa ser modelada através da regressão logística. Tudo isso se deve, principalmente, pela facilidade de interpretação dos parâmetros de um modelo logístico e também pela possibilidade do uso desse tipo de metodologia em problemas de classificação. A questão de interpretação dos parâmetros é crucial pois muitos modelos ou métodos utilizados no problema da classificação são de difícil interpretação.

A regressão logística é freqüentemente apropriada para a análise de experimentos que apresentam variáveis resposta categóricas em que o interesse seja o de descrever a relação entre a variável resposta e um conjunto de variáveis explanatórias (covariáveis) (ver [10]). Quando a variável resposta é dicotômica (somente duas categorias), tem-se a, assim denominada, regressão logística dicotômica. Para variáveis resposta com mais que duas categorias a denominação usada é regressão logística politômica. As covariáveis, em regressão logística, podem ser categóricas ou contínuas. Variáveis *dummies* são usadas para que as covariáveis categóricas sejam consideradas em um modelo de regressão logística.

Conforme [28], na pesquisa médica estamos freqüentemente interessados na ocorrência de um certo evento, normalmente associado com a sobrevivência do paciente, e os fatores que influenciam esta ocorrência. A análise estatística, usualmente necessária, consiste em relacionar, através de um modelo a variável resposta com os fatores de risco que foram medidos. Quando a resposta é dicotômica, o que se procura é um modelo que relacione a probabilidade de ocorrência de um evento. É comum na literatura, usar-se 1 para representar a resposta de maior interesse, e chamá-la de *sucesso* e 0 para a outra resposta, chamada normalmente de *falha*.

A Função Logística

No contexto apresentado em [11] a função logística definida por:

$$f(X) = \frac{\alpha}{1 + e^{-(\beta+\gamma X)}} \tag{13}$$

Nesta formulação, α, β e γ são parâmetros, $\alpha > 0$ e $\gamma > 0$, foi indicada para o estudo descritivo do crescimento de populações humanas por Verhulst (1845), que denominou de "curva logística".

Muitos anos mais tarde, Pearl e Reed (1920), sem conhecerem a contribuição de Verhulst (1845), obtiveram a mesma curva, que utilizaram para descrever o crescimento da população dos EUA, de 1790 a 1910, com base em dados censitários.

A partir daí, a curva logística tem sido bastante estudada quanto às suas características matemáticas e quanto ao método de estimar às suas características matemáticas e quanto ao método de estimar seus parâmetros. Ela também tem sido largamente empregada para a representação de dados empíricos de crescimento de animais e vegetais, de

crescimento de populações humanas e de adoção de novos bens econômicos.

$$\frac{d}{dX}f(X) = \frac{\alpha\gamma e^{-(\beta+\gamma X)}}{[1+e^{-(\beta+\gamma X)}]^2} = \frac{\gamma}{\alpha}f(X)[\alpha - f(X)] \tag{14}$$

temos:

$$\lim_{x\to\infty} f(X) = \alpha \tag{15}$$

e

$$\lim_{x\to\infty} f(X) = 0 \tag{16}$$

Verifica-se que a função logística é monotonicamente crescente e fica entre duas assíntotas horizontais que são o eixo das abcissas e a reta de ordenadas constante igual a α. O parâmetro α, que é a distância entre as duas assíntotas, é denominado "nível de saturação". O parâmetro γ está relacionado com a taxa de crescimento da função. E, finalmente, β é um parâmetro de posição ou locação, isto é, mudando o valor de β enquanto todos os outros parâmetros são mantidos fixos, a curva apenas se movimenta horizontalmente.

A Função Distribuição Logística

Para completar a teoria apresentada por HOFFMANN 1977 e já referenciada neste trabalho, vamos utilizar muitos elementos apresentados em GIOLO 2006, para variáveis aleatórias binárias.

Uma diferença importante entre o modelo de regressão logística e o modelo de regressão linear pode ser notada e, esta, diz respeito à natureza da relação entre a variável resposta e as variáveis independentes. Em qualquer problema de regressão a quantidade sendo modelada é o valor médio da variável resposta dado os valores das variáveis independentes. Esta quantidade é chamada média condicional e será expressa por $E(Y|x)$, em que Y denota a variável resposta e x denota os valores das variáveis independentes. Em regressão linear tem-se $-\infty < E(Y|x) < +\infty$ e, em regressão logística, devido à natureza da variável resposta, $0 \leq E(Y|x) \leq 1$. Observe que a mudança em $E(Y|x)$ por unidade de mudança em x torna-se progressivamente menor quando $E(Y|x)$ torna-se próxima de zero ou de um. A curva em forma de "S" lembra a distribuição acumulada de uma variável aleatória, o que motivou o uso da distribuição logística para fornecer um modelo para $E(Y|x)$.

A função de distribuição logística é descrita por:

$$F(x) = \frac{1}{1 + exp\{-x\}} = \frac{exp\{x\}}{1 + exp\{x\}} \tag{17}$$

em que, para $x = -\infty$ e $x = +\infty$, tem-se $F(-\infty) = 0$ e $F(+\infty) = 1$.

A função de distribuição logística toma valores entre zero e um; assume o valor zero em uma parte do domínio das variáveis explicativas, um em outra parte do domínio e cresce suavemente na parte intermediária possuindo uma particular curva em forma de "S". Outras funções de distribuição possuem as características mencionadas. No entanto, a função logística foi escolhida basicamente por duas razões.

1. do ponto de vista matemático é extremamente flexível e fácil de ser usada e,
2. conduz a interpretações simples.

Para descrever a variação entre os $\theta(x)$= $E(Y|x)$, foi, então, proposto o modelo de regressão logística expresso por:

$$\theta(x) = P(Y = 1|x) = \frac{exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right\}}{1 + exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right\}} \tag{18}$$

em que $Y_i = 1$ significa a presença da característica de interesse na variável resposta, x representa as covariáveis [2], isto é, $x = (x_1, x_2, \ldots, x_p)$, o parâmetro β_0 é o intercepto e β_k $(k = 1, \ldots, p)$ são os p parâmetros de regressão. Observe que este modelo retornará uma estimativa da probabilidade do indivíduo apresentar a resposta de interesse dado que o mesmo possui, ou não, determinadas características que estão expressas nas covariáveis.

Conseqüentemente,

$$1 - \theta(x) = \frac{exp\left\{-(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)\right\}}{1 + exp\left\{(-\beta_0 + \sum_{k=1}^{p} \beta_k x_k)\right\}} \tag{19}$$

$$= \frac{1}{1 + exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right\}} \tag{20}$$

retornará uma estimativa da probabilidade do indivíduo não apresentar a resposta de interesse dado suas informações mensuradas na covariável

[2]Estes são comumente os fatores de risco em problemas ligados à área médica.

Função de Ligação Logito

Observe, ainda, que a função:

$$log(\frac{\theta(x)}{1-\theta(x)}) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k \tag{21}$$

em θ é linearmente relacionada com as covariáveis. Esta função é chamada de função **logito**, isto é, o logaritmo neperiano da razão entre $\theta(x)$ e 1 - $\theta(x)$. O *logito* é, na realidade, o logaritmo de uma razão de chances (*odds*) e, este fato, permitirá que razões de chance (*odds ratios*) sejam, portanto, derivadas do modelo.

No contexto de modelos lineares generalizados, uma função, monótona e derivável, que relaciona a média ao preditor linear é denominada de função de ligação. Assim, $\eta = log(\frac{\theta(x)}{1-\theta(x)})$, é a função de ligação canônica para o modelo binomial.

É importante ressaltar que o modelo prediz probabilidades, ou seja, números entre 0 e 1, não sendo possível encontrar valores fora deste domínio.

Outra diferença importante entre um modelo clássico de regressão linear e o modelo de regressão logística refere-se à distribuição condicional da variável resposta. Quando a resposta é dicotômica ($Y = 1$ ou 0), o valor da variável resposta dado x é expresso por $y = \theta(|x) + \varepsilon$ e, como a quantidade ε pode assumir somente um de dois possíveis valores, isto é, $\varepsilon = 1 - \theta(x)$ para $y = 1$ ou, $\varepsilon = -\theta(x)$ para $y = 0$, segue que ε tem distribuição com média zero e variância dada por $\theta(x)(1-\theta(x))$, isto é, a distribuição condicional da variável resposta segue uma distribuição Binomial com probabilidade dada pela média condicional $\theta(x)$ (ver GIOLO 2006).

Estimadores de Máxima Verossimilhança

A estimação dos parâmetros em regressão logística é feita, em geral, pelo método de máxima verossimilhança. Para aplicação deste método é necessário, inicialmente, construir a função de verossimilhança, a qual expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores de máxima verossimilhança dos parâmetros serão os valores que maximizam esta função.

Seja dada uma amostra de n valores de X_i e Y_i. Se os u_i são variáveis aleatórias independentes com distribuição normal de média zero e variância σ^2, a função de verossimilhança da amostra é[11]

$$L = \prod_{i=1}^{n} (2\pi\sigma^2)^{\frac{-1}{2}} exp\left\{-\frac{(Y_i - \alpha - \beta\rho^{x_i})^2}{2\sigma^2}\right\} \tag{22}$$

As estimativas de máxima verossimilhança ($\hat{\alpha}$, $\hat{\beta}$, $\hat{\rho}$ e $\hat{\sigma}^2$) são os valores de α, β, ρ e σ^2 que maximizam L e, portanto, também maximizam l_nL. Conforme apresentado em HOFFMANN (1977) Temos:

$$l_nL = -\frac{n}{2}l_n2\pi - \frac{n}{2}l_n\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta\rho^{x_i})^2 \tag{23}$$

Igualando a zero as derivadas parciais de l_nL em relação a α, β, ρ e σ^2, obtemos as seguintes equações:

$$\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}\hat{\rho}^{x_i}) = 0 \tag{24}$$

$$\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}\hat{\rho}^{x_i})\hat{\rho}^{x_i} = 0 \tag{25}$$

$$\hat{\beta}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}\hat{\rho}^{x_i})X_i\hat{\rho}^{x_i-1} = 0 \tag{26}$$

e

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}\hat{\rho}^{x_i})^2 \tag{27}$$

As três primeiras equações constituem um sistema com $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\rho}$ como incógnitas. Comparando esse sistema com o sistema de mínimos quadrados, concluímos que as estimativas de máxima verossimilhança para α, β e ρ coincidem com as estimativas de mínimos quadrados (a, b, r) cujas expressões são encontradas em HOFFMANN(1977). Então, o estimador de máxima verossimilhança da variância do erro é:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - a - br^{x_i})^2 \tag{28}$$

Estimadores não tendenciosos para a variância do erro são encontrados pela divisão da soma dos quadrados dos desvios por n-p (p é o número de parâmetros da equação). A forma deste estimador é, neste caso:

$$s^2 = \frac{1}{n-3}\sum_{i=1}^{n}(Y_i - a - br^{x_i})^2 \tag{29}$$

No modelo de regressão logística, considerando a variável resposta Y codificada como zero ou um , pode-se obter a probabilidade condicional de que Y seja igual a 1 dado x, isto é, $\theta(x) = P(Y = 1|x)$ e, em conseqüência, a probabilidade condicional de que Y seja igual a zero dado x, isto é, 1 - $\theta(x) = P(Y = 0|x)$. Assim, $\theta(x_i)$ será a contribuição para a função de verossimilhança dos pares (y_i,x_i) em que $y_i = 1$ e $1-\theta(x_i)$, a contribuição dos pares em que $y_i = 0$.

Assumindo-se que as observações são oriundas de variáveis independentes, tem-se a seguinte expressão para a função de verossimilhança:

$$L(\beta) = \prod_{i=1}^{n} (\theta(x_i))^{y_i} (1 - \theta(x_i)^{1-y_i} \tag{30}$$

As estimativas de β serão os valores que maximizam a função de verossimilhança. Algebricamente é mais fácil trabalhar com o logaritmo desta função, isto é, com:

$$l(\beta) = logL(\beta) = \sum_{i=1}^{n} y_i log(\theta(x_i)) + (1 - y_i) log(1 - \theta(x_i)) \tag{31}$$

Para obter os valores de *beta* que maximixam (β) basta diferenciar a respectiva função com respeito a cada parâmetro β_j (j = 0, 1, 2, ..., p) obtendo-se, assim, o sistema de $p+1$ equações, que quando igualadas a zero, produzem como solução as estimativas de máxima verossimilhança de β conforme em GIOLO2006.

Os valores ajustados para o modelo de regressão logística são, portanto, obtidos substituindo-se as estimativas de β na expressão abaixo:

$$\theta(x) = P(Y = 1|x) = \frac{exp\left\{\beta_0 + \sum\limits_{k=1}^{p} \beta_k x_k\right\}}{1 + exp\left\{\beta_0 + \sum\limits_{k=1}^{p} \beta_k x_k\right\}} \tag{32}$$

As $p + 1$ equações são chamadas equações de verossimilhança e por serem não-lineares nos parâmetros j (j = 0, 1,.., p), requerem métodos especiais para encontrar suas soluções. Os métodos iterativos de Newton-Raphson e o escore de Fisher são algoritmos numéricos comumente utilizados com esta finalidade.

Comentários sobre o uso da EMV

Vamos apresentar algumas considerações sobre os estimadores de máxima verossimilhança que podem ser vistas em [22].

O estimador de máxima verossimilhança tem muitas propriedades assintóticas que tornam o seu resultado mais atrativo. O estimador é assintoticamente consistente, significando que o aumento no tamanho da amostra, aproxima os valores das estimativas dos valores verdadeiros. Ele é não-viesado, isto é, sua a esperança é igual ao valor estimado. Ele também é assintoticamente eficiente, quanto maior a amostra, maior precisão das estimativas. Estes estimadores são normalmente distribuídos. Estas são excelentes propriedades da teoria das grandes amostras.

Uma das desvantagens dos estimadores de máxima verossimilhança é a necessidade de grandes amostras para garantir as propriedades ótimas: 30 a 50 observações, dependendo da aplicação. Com poucos dados, o método pode criar viés. Sabe-se, por exemplo, que as estimativas de MV para o parâmetro de forma (beta) da distribuição Weibull são viesadas para tamanhos de amostra pequenos, e o efeito pode ser aumentado dependendo da quantidade de censura. Esta tendência pode causar discrepância na análise.

Há também situações mórbidas quando as situações assintóticas do MLE não se aplicam. Um destas é a estimação do parâmetro de escala (eta) da distribuição Weibull três parâmetros quando o parâmetro de forma (beta) tem um valor perto de 1. Esses problemas, podem causar maiores discrepâncias.

Em geral, é recomendado usar a técnica regressão em X, quando se tem uma amostra pequena e sem censuras. Quando se tem presente, censuras em grande quantidade ou de forma desigual as falhas, quando uma proporção grande de dados está presente ou o tamanho da amostra é suficiente, é o estimador de máxima verossimilhança deve ser escolhido.

Critérios para Escolha de Modelos

Após a obtenção das estimativas dos coeficientes β_j (j=0,1,...,20) avaliamos a adequação do modelo ajustado. Nosso interesse é comparar os valores observados da variável resposta com os valores preditos pelo modelo, testando a significância das covariáveis no modelo.

Conforme [10], a comparação pode ser feita utilizando-se o teste da razão de verossimilhanças, em que a função de verossimilhança do modelo sem as covariáveis L_{SC} é comparada com a função de verossimilhança do modelo com as covariáveis L_{CC}. Formalmente, o teste é expresso por:

$$\begin{aligned} RV &= -2log\left[\frac{verossimilhança\ do\ modelo\ sem\ covariáveis}{verossimilhança\ do\ modelo\ com\ as\ covariáveis}\right] \\ &= -2log\left[\frac{L_{SC}}{L_{CC}}\right] \\ &= 2log(L_{CC}) - 2log(L_{SC}) \end{aligned}$$

Note que a razão das verossimilhanças é multiplicada por -2log. Isto é feito para que se obtenha uma quantidade cuja a distribuição é conhecida (no caso a distribuição qui-quadrado) de modo que, tal quantidade, possa ser usada para a realização de testes de hipóteses. Em regressão logística a estatística, abaixo, é chamada **deviance**.

$$D = -2log\left[\frac{verossimilhança\ do\ modelo\ sob\ estudo}{verossimilhança\ do\ modelo\ saturado}\right]$$

Um modelo saturado é aquele que contém tantos parâmetros quantos dados existirem. Assim a estatística RV pode ser vista como a diferença de duas *deviances*. Sob a hipótese nula de que os p coeficientes associados às covariáveis no modelo são iguais a zero, a distribuição de RV será qui-quadrado com p graus de liberdade. Rejeição da hipótese nula, neste caso tem interpretação análoga àquela em regressão linear, ou seja, pode-se concluir que pelo menos um, ou talvez todos os p coeficientes, sejam diferentes de zero (ver [10]).

Utilizaremos a *deviance* como uma medida de discrepância do modelo, considerando o experimento com as vinte covariáveis e a partir das *deviances* e suas diferenças, através do teste de razão de verossimilhança, poderemos testar a significância da inclusão e/ou interações de determinadas covariáveis no modelo.

É necessário também observar a correlação existente entre as covariáveis.

2.1.4 ANÁLISE LINEAR DISCRIMINANTE

Segundo HAIR (2006), a análise discriminante envolve determinar uma variável resultante da combinação linear das duas (ou mais) variáveis independentes que discriminarão melhor entre grupos definidos a priori. A discriminação é conseguida estabelecendo-se os pesos de cada variável que resultem na maximização da variância entre grupos relativa à variância dentro dos grupos. A combinação linear para uma análise discriminante, também conhecida como função discriminante, é determinada de uma equação que assume a seguinte forma:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + ... + +W_nX_{nk} \tag{33}$$

Onde:
Z_{jk} = escore Z discriminante da função discriminante j para o objeto k
a = intercepto
W_j = peso discriminante para a variável independente i
X_{ik} = variável independente i para o objeto k

A análise discriminante é a técnica estatística apropriada para testar a hipótese de que as médias de um grupo de um conjunto de variáveis independentes para dois ou mais grupos são iguais. Para tanto, a análise discriminante multiplica cada variável independente por seu peso correspondente e acrescenta esses produtos juntos. O resultado é um escore Z discriminante composto para cada indivíduo na análise. Calculando a média dos escores discriminantes para todos os indivíduos em um grupo, conseguimos a média do grupo. Essa média de grupo é chamada centróide. Quando a análise envolve dois grupos, há dois centróides; com três grupos, há três centróides e assim por diante. Os centróides indicam o local mais típico de qualquer indivíduo de um grupo particular, e uma comparação dos centróides de grupos mostra o quão afastados estão os grupos ao longo da dimensão testada.

O teste para a significância estatística da função discriminante é uma medida generalizada da distância entre os centróides de grupos. Ela é computada comparando as distribuições dos escores discriminantes para os grupos. Se a sobreposição nas distribuições é pequena, a função discriminante separa bem os grupos. Se a sobreposição é grande, a função é um discriminador pobre entre os grupos (ver HAIR, 2006).

A análise discriminante múltipla é uma extensão em que há mais de dois grupos conhecidos a priori na variável dependente, resultando portanto no cálculo de mais de uma função discriminante. De fato, calculam-se $G^{\smile}1$ funções, onde G é o número de grupos. Cada função discriminante calcula um escore discriminante Z. No caso de uma variável dependente de três grupos, cada objeto terá um escore para funções discriminantes um e dois, permitindo que os objetos sejam representados graficamente em duas dimensões, com cada dimensão representando uma função discriminante. Logo, a análise discriminante não está limitada a uma única variável, como ocorre na regressão múltipla, mas cria múltiplas variáveis que representam dimensões de discriminação entre grupos.

OBJETIVOS DA ANÁLISE DISCRIMINANTE

A análise discriminante pode abordar qualquer um dos seguintes objetivos de pesquisa:

1. Determinar se existem diferenças estatisticamente significantes entre os perfis de escore médio em um conjunto de variáveis para dois (ou mais) grupos definidos a priori.
2. Determinar quais das variáveis independentes explicam o máximo de diferenças nos perfis de escore médio dos dois ou mais grupos.
3. Estabelecer procedimentos para classificar objetos (indivíduos, produtos e assim por diante) em grupos, com base em seus escores em um conjunto de variáveis independentes.
4. Estabelecer o número e a composição das dimensões de discriminação entre grupos formados a partir do conjunto de variáveis independentes.

Como pode ser notado a partir desses objetivos, a análise discriminante é útil quando o pesquisador está interessado em compreender diferenças de grupos ou em classificar objetos corretamente em grupos ou classes. Portanto, a análise discriminante pode ser considerada um tipo de análise de perfil ou uma técnica preditiva analítica. Em qualquer caso, a técnica é mais apropriada onde existe uma só variável dependente categórica e diversas variáveis independentes métricas. Como uma análise de perfil, a análise discriminante fornece uma avaliação objetiva de diferenças entre grupos em um conjunto de variáveis independentes (HAIR, 2006).

SELEÇÃO DE VARIÁVEIS DEPENDENTE E INDEPENDENTES

Para aplicar a análise discriminante, o pesquisador deve primeiramente especificar quais variáveis devem ser independentes e qual deve ser a dependente. Lembrando que a variável dependente é categórica e as independentes são métricas (contínuas).

FUNÇÃO DISCRIMINANTE LINEAR DE FISHER

Basicamente, o problema consiste em separar duas classes de objetos ou fixar um novo objeto em uma das duas classes. É comum denominar as populações de π_1 e π_2 e os objetos separados ou classificados com base nas medidas de p variáveis aleatórias são associadas com vetores do tipo:

$$X' = [X_1, X_2, ..., X_p]$$

em que as variáveis $X_i, i = 1, 2, ..., p$, são as medidas das características investigadas nos objetos. Os valores observados de X podem diferir de uma classe para outra, sendo que a totalidade dos valores da 1^{a} classe é a população dos valores X para π_1 e aqueles da 2^{a} classe a população são a população dos valores de X para π_2. Assim, estas populações podem ser descritas pelas funções densidade de probabilidade $f_1(x)$ e $f_2(x)$ conforme ANSELMO (2006).

Pela função discriminante de Fisher observações multivariadas da matriz X são transformadas em univariadas Y, de forma que Y informa sobre as populações π_1 e π_2. Se estas populações forem as mais distintas o quanto for possível, fica mais fácil afirmar a qual delas pertence uma determinada observação; mas isto nem sempre acontece e as populações ocupam algumas áreas em comum no espaço, denominadas "regiões de confusão". Para resolver esse problema, Fisher, em 1936, sugeriu tomar a combinação linear de X para criar Y (y=l'x), por ser uma função simples de X e de fácil tratamento matemático. Tendo $\mu 1y(E(l'x/\pi 1))$ como a média dos resultados Y, obtida das X, cujas observações pertencem a π_1 e $\mu 2y$ $(E(l'x/\pi_2))$ a média de Y obtida de X que pertence a π_2, Fisher selecionou a combinação linear que maximiza o quadrado da distância entre $\mu 1y$ e $\mu 2y$ relativa à variabilidade de X nas duas populações, dadas pelas matrizes de covariância

$\Sigma = Ei[(x - \mu_i)(x - \mu_i)']$, i = 1, 2, considerada igual para duas populações. Nessa matriz, μ_1 e μ_2 são, respectivamente, a média da população de X da população π_1 e média de X da população π_2. A distância máxima das duas populações é dada por

$$(x - \mu_1)'\Sigma - 1(x - \mu_2)$$

. Naturalmente as quantidades populacionais μ_1, μ_2 e Σ raramente são conhecidas e a expressão anterior só poderá ser utilizada se forem estimadas as quantidades populacionais (ver TSUCHIYA, 2002).

Têm-se n_1 observações da variável multivariada $X' = [x1, x2, ..., xp]$ de π_1 e n_2 medidas dessa quantidade de π_2. Sejam as seguintes estatísticas relativas as amostras, denotando, respectivamente, a média amostral e variância amostral:

$$\bar{x} = \sum_{j=1}^{n_i} X_{ij} \tag{34}$$

$$S_i = \frac{1}{n_{i-1}} \sum_{j=1}^{n_i} (X_{ij} - \bar{x})(X_{ij} - \bar{x})' \tag{35}$$

onde i= 1, 2

A função discriminante de Fisher é construída sem assumir a existência de uma função de probabilidade associada a cada grupo. Fisher propõe uma função linear $y = l'x$, que maximiza a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos. Porém, se as duas populações têm uma matriz de variância e covariância comum, e isto é muito conveniente sob o ponto de vista matemático, a matriz S pode ser substituída pela matriz S_{pooled} (combinado):

$$S_{pooled} = \frac{(n_1 - 1)}{(n_1 - 1)(n_2 - 1)} S_1 + \frac{(n_2 - 1)}{(n_1 - 1)(n_2 - 1)} S_2 \tag{36}$$

Para alocar o objeto na população m, primeiramente há a necessidade de definir o ponto médio da combinação linear, ou seja,

$$m = \frac{1}{2}(\bar{x}_1 - \bar{x}_2) = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}_{pooled}(\bar{x}_1 - \bar{x}_2) \tag{37}$$

Então uma observação x_0 será classificada como pertencente à população π_1 se:

$$(\bar{x}_1 - \bar{x}_2)' S^{-1}_{pooled} x_0 \geq m \tag{38}$$

e alocada para o grupo π_2 se:

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0 \leq m \tag{39}$$

Quando p=2 o conjunto de dados é tão discriminado se projetado nos eixos $x_1 e x_2$, assim sendo, o método rebate os dados numa função linear dos dois eixos. A melhor decisão é a que torna máxima a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos conforme já comentado..

Segundo JOHNSON & WICHERN (1992), este tipo de análise só faz sentido se as duas populações realmente tiverem médias diferentes. Suponha que as populações π_1 e π_2 sejam normais multivariadas com uma matriz de covariância comum Σ. Um teste de $H_0 : \mu_1 = \mu_2$, contra $H_1 : \mu_1 \neq \mu_2$ é feito pela estatística $(n1 + n2 - p - 1)/[(n1 + n2 - 2)p)](n1n2)/(n1 +$ que tem distribuição F com $v_1 = p$ e $v_2 = n_1 + n_2 - p - 1$ graus de liberdade em que:

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2). \tag{40}$$

Se H_0 for rejeitada, pode-se concluir que a separação entre as duas populações π_1 e π_2 é significativa, caso contrário não há evidências contra a hipótese de que as duas populações têm a mesma média e matriz de covariância, ou seja, elas formam uma única população.

2.1.5 REDES NEURAIS ARTIFICIAIS

O cérebro humano é considerado o mais fascinante processador baseado em carbono existente, sendo composto por aproximadamente 10 bilhões neurônios. Todas as funções e movimentos do organismo estão relacionados ao funcionamento destas pequenas células. Os neurônios estão conectados uns aos outros através de sinapses, e juntos formam uma grande rede, chamada REDE NEURAL. As sinapses transmitem estímulos através de diferentes concentrações de Na+ (Sódio) e K+ (Potássio), e o resultado disto pode ser estendido por todo o corpo humano. Esta grande rede proporciona uma fabulosa capacidade de processamento e armazenamento de informação.

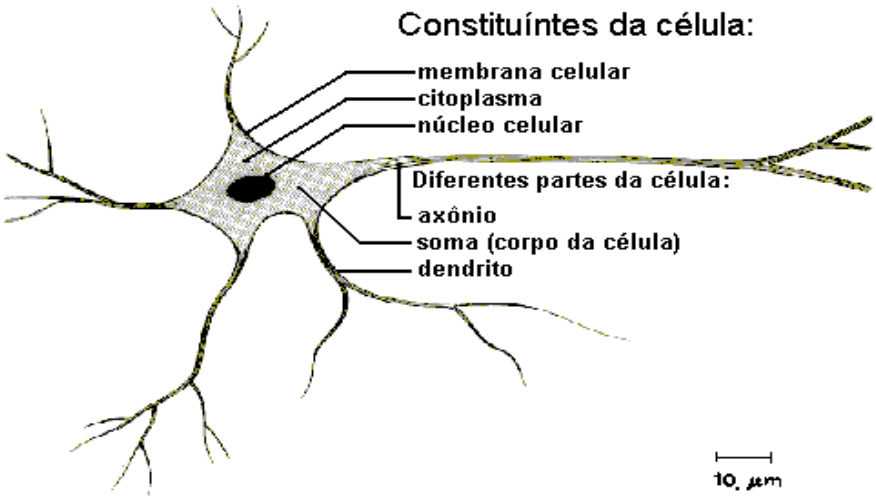
O sistema nervoso é formado por um conjunto extremamente complexo de neurônios. Nos neurônios a comunicação é realizada através de impulsos, quando um impulso é recebido, o neurônio o processa, e passado um limite de ação, dispara um segundo impulso que produz uma substância neurotransmissora o qual flui do corpo celular para o axônio (que por sua vez pode ou não estar conectado a um dendrito de outra célula). O neurônio que transmite o pulso pode controlar a freqüência de pulsos aumentando ou diminuindo a polaridade na membrana pós sináptica. Eles tem um papel essencial na determinação do funcionamento, comportamento e do raciocínio do ser humano. Ao contrário das redes neurais artificiais, redes neurais naturais não transmitem sinais negativos, sua ativação é medida pela freqüência com que emite pulsos, freqüência esta de pulsos contínuos e positivos. As redes naturais não são uniformes como as redes artificiais, e apresentam

uniformidade apenas em alguns pontos do organismo. Seus pulsos não são síncronos ou assíncronos, devido ao fato de não serem contínuos, o que a difere de redes artificiais (informações em [27].

Conforme pode ser observado na figura 1 as principais componentes dos neurônios são:

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

Figura 1: Componentes do Neurônio



Histórico

Após a publicação em 1986 do hoje clássico Parallel Distributed Processing, editado por Rummelhart e McClelland doPDP Research Group da Universidade da Califórnia em San Diego, a área de redes neurais teve um desenvolvimento explosivo com a multiplicação exponencial de journal's, associações locais e internacionais, sem falar da torrente de teses e paper's científicos. No começo desta década surgiram várias empresas para exploração comercial de produtos de redes neurais, invariavelmente produtos de software para o desenvolvimento de aplicações ou simulações acadêmicas.

NEUROCOMPUTAÇÃO

Os modelos neurais procuram aproximar o processamento dos computadores ao cérebro. As redes neurais possuem um grau de interconexão similar à estrutura do cérebro. Em computador convencional moderno a informação é transferida em tempos específicos dentro de um relacionamento com um sinal para sincronização.

A tabela 3, traça um comparativo entre o cérebro humano e o computador:

Parâmetro	Cérebro	Computador
Material	Orgânico	Metal e Plastico
Velocidade	Milisegundos	Nanosegundos
Tipo de Processamento	Paralelo	Sequencial
Armazenamento	Adaptativo	Estático
Controle de Processos	Distribuído	Centralizado
Número de elementos Processados	10 e 11 à 10 e 14	10 e 5 à 10 e 6
Ligações entre elementos processados	10.000	<10

Tabela 3: Comparativo entre o cérebro humano e o computador

O mesmo paralelo pode ser traçado comparando o computador com as redes neurais, conforme apresentado na tabela 4. Para tanto, a comparação não se dará com um computador específico encontrado no mercado, mas sim com o paradigma predominante nos computadores atuais.

Computadores	Neurocomputadores
Executa Programas	Aprende
Executa Operações lógicas	Executa informações não lógicas, transformações, comparações
Depende do modelo ou do programador	Descobre as relações ou regras dos dados e exemplos
Testa uma hipótese por vez	Testa todas as possibilidades em paralelo

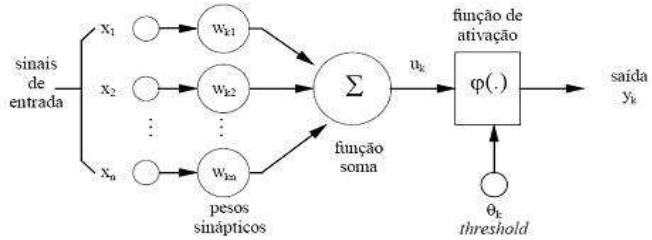
Tabela 4: Comparativo entre o computador e as redes neurais

O NEURÔNIO ARTIFICIAL

O neurônio artificial é uma unidade de processamento (ou célula de rede) matematicamente simples, que recebe uma ou mais entradas e as transforma em saídas. Cada entrada tem um peso associado, que determina sua intensidade. A Figura a seguir, mostra um esquema de neurônio artificial.[2]

Conforme mencionado em [2], é possível distinguir alguns elementos importantes no neurônio mostrado na Figura 2:

Figura 2: Neuronio Artificial



1. as **sinapses**, que são caracterizadas por um peso, w, que pode representar a sua intensidade. O papel do peso wkj é multiplicar o sinal xj na entrada da sinapse j, conectada a um neurônio k. O peso wkj é positivo se a sinapse associada é excitatória e negativo se a sinapse associada é inibitória;

2. um **somatório**, que adiciona as entradas ponderadas pelos seus pesos respectivos, ou seja,
$$u_k = \sum_{j=1}^{n} W_{kj} X_j$$

3. um **limiar** (threshold), θ_k, que tem um papel determinante na saída do neurônio. Se o valor de u_k for menor que este limiar, então, a saída do neurônio fica inibida. Caso contrário, o neurônio fica ativo.

4. uma **função de ativação**, que funciona como um limitante à amplitude da saída do neurônio, ou seja, a entrada é normalizada dentro de um intervalo fechado, comumente [0,1] ou [-1,1];

5. a **saída** do neurônio, y_k, onde: $y_k = \Phi(u_k - \theta_k)$ onde Φ é a função de ativação.

Em geral, o valor do *threshold* é aplicado com a inclusão de uma entrada x_0 igual a -1 e um peso w_{k0} igual ao valor de θ_k. Portanto, a nova entrada da função de ativação, já incluindo o limiar, é dada por:

$$\nu_k = \sum_{j=1}^{n} W_{kj} X_j - \theta_k$$

TIPOS DE FUNÇÃO DE ATIVAÇÃO

A função de ativação define a saída do neurônio em termos do nível de atividade do mesmo. Dentre as funções de ativação mais comuns, se enquadram as seguintes:

- função linear:

$$\Phi_i(t+1) = v_i t \tag{41}$$

- função threshold ou limiar:

$$\Phi_i(t+1) = 1, sev_i t \geq \theta \Phi_i(t+1) = 0, sev_i t < \theta \tag{42}$$

- função sigmóide logística:

$$\Phi_i(t+1) = \frac{1}{(1+e^{-vi(t)})} \tag{43}$$

- função tangente hiperbólica:

$$\Phi_i(t+1) = \frac{(1-e^{-vi(t)})}{(1+e^{-vi(t)})} \tag{44}$$

- função linear por partes:

$$\Phi_i(t+1) = +1, sev_i t > \theta \Phi_i(t+1) = -1, sev_i t < \theta \tag{45}$$

De forma geral, conforme apresentado em [27] a operação de uma célula da rede se resume em:

1. Sinais são apresentados à entrada;
2. Cada sinal é multiplicado por um peso que indica sua influência na saída da unidade;
3. É feita a soma ponderada dos sinais que produz um nível de atividade;
4. Se este nível excede um limite (threshold) a unidade produz uma saída.

As redes Neurais Artificiais

As Redes Neurais artificiais são baseadas em modelos abstratos do funcionamento do cérebro humano e tentam reproduzir sistemas biologicamente realísticos.(ver [2])

Informalmente uma rede neural artificial (RNA) é um sistema composto por vários neurônios. Conforme citado em [1] estes neurônios estão ligados por conexões, chamadas conexões sinápticas. Alguns neurônios recebem excitações do exterior e são

chamados neurônios de entrada e correspondem aos neurônios dos órgãos dos sentidos. Outros têm suas respostas usadas para alterar, de alguma forma, o mundo exterior e são chamados neurônios de saída e correspondem aos motoneurônios que são os neurônios biológicos que excitam os músculos. Os neurônios que não são nem entrada nem saída são conhecidos como neurônios internos. Estes neurônios internos a rede tem grande importância e são conhecidos na literatura saxônica como "hidden" fazendo com que alguns traduzam como "escondidos". Os neurônios internos são importantes por vários aspectos:

- Importância biológica: Por corresponder a uma atividade do sistema nervoso que pode apresentar uma independência de excitações externas. Com efeito, se entre estes neurônios houver ligações formando ciclos, e considerando ainda um certo tempo de resposta de um neurônio, após cessar toda excitação exterior pode haver nestes neurônios internos uma evolução de um vetor representativo da excitação destes neurônios. Esta excitação pode provocar uma evolução durante um tempo relativamente longo e pode ser interpretada como uma metáfora da mente, onde pensamentos vêm e voltam, sem estímulo exterior.
- Importância matemática: Desde que se provou que sem estes neurônios é impossível uma RNA resolver problemas classificados como linearmente não separáveis.

[2]cita alguns benefícios da computação neural, dentre eles:

- **habilidade de aprender com exemplos**: os computadores neurais têm a capacidade de aprender com a experiência, objetivando melhorar seu desempenho e se adaptar a ambientes novos e dinâmicos, diferentemente dos computadores comuns;
- **robustez**: as Redes Neurais tem habilidade em lidar com ruídos. Elas são tolerantes à falhas e podem apresentar degradação gradual, ou seja, apesar de alguma falha no sistema, elas continuam a fornecer respostas adequadas por um bom tempo, o que as diferencia dos computadores convencionais, onde uma falha pode causar prejuízo do sistema como um todo;
- **velocidade de processamento**: como as Redes Neurais consistem de um grande número de unidades de processamento operando em paralelo, elas podem operar em velocidades consideráveis em relação aos métodos computacionais comuns.

Além das vantagens já mencionadas, em muitos casos e dependendo do problema para o qual são utilizadas, o desempenho das Redes Neurais têm se mostrado bastante superior aos métodos estatísticos convencionais, usados para o mesmo fim.

CARACTERIZAÇÃO DAS RNA'S

Para caracterizar uma RNA é importante especificar os seguintes pontos (ver [1]):

- Os componentes da rede: os neurônios: ex: estáticos? Dinâmicos?
- A resposta de cada neurônio: dicotômica? Intervalo dos reais?
- O estado global de ativação da rede: vetor cujas componentes são as ativações dos neurônios?
- A conectividade da rede dada pelos valores de conexões sinápticas: que define a topologia da rede.
- Como se propaga a atividade da rede: síncrona? assíncrona?
- Como se estabelece a conectividade da rede: aprendizado.
- O ambiente externo à rede: estático? Dinâmico? Aleatório? Determinístico?
- Como o conhecimento é representado na rede: localizado? Distribuído?

Resumindo, as RNA podem ser categorizadas por sua topologia, isto é, pelo número de camadas, de elementos de processamento e de conexões; pelas características de seus elementos de processamento; e pelas leis de aprendizagem a que foram submetidas conforme citado em [18].

TOPOLOGIA DAS RNA'S

A maneira como os neurônios são organizados é chamada de topologia da rede. A topologia irá afetar o desempenho da rede ([2]), assim como as aplicações para as quais ela é desejada, e sua estrutura está intimamente ligada ao algoritmo de aprendizado usado para a fase de treinamento. Algumas redes permitem que as conexões caminhem tanto no sentido entrada-saída, quanto saída-entrada. Outras permitem que os neurônios da mesma camada estejam conectados. Ainda há as que permitem que o neurônio envie sinais de volta para ele mesmo (retroalimentação).

Dentre as topologias, pode-se citar:

- **Multlayer Perceptron**: é formada de uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída e é o tipo de arquitetura mais encontrado.
- **Kohonen**: é formada por uma camada de entrada e uma única camada de saída, onde cada neurônio está conectado a todos os seus vizinhos.
- **Hopfield**: não há neurônios de entrada ou saída. A entrada de um neurônio é a saída dos demais neurônios e a sua saída vai ser a entrada do demais neurônios da rede.
- **ART**: é formada por camada de entrada e saída, além de controladores de fluxo de sinais.

APRENDIZADO

Conforme [1], aprender é o ato que produz um comportamento diferente a um estímulo externo devido à excitações recebidas no passado e é de uma certa forma sinônimo de aquisição de conhecimento. RNA possuem a capacidade de aprenderem por exemplos, e fazerem interpolações do que aprenderam.

O aprendizado de uma RNA pode se supervisionado, não supervisionado ou híbrido.

No aprendizado supervisionado são sucessivamente apresentadas à rede conjuntos de padrões de entrada e seus correspondentes padrões de saída. Em [18], durante este processo, a rede realiza um ajustamento dos pesos das conexões entre os elementos de processamento, segundo uma determinada lei de aprendizagem, até que o erro entre os padrões de saída gerados pela rede alcancem um valor mínimo desejado. Por exemplo, *perceptron, adaline e madaline, backpropagation*, são algumas dentre as dezenas de leis de aprendizagem supervisionada.

No aprendizado não-supervisionado a rede "analisa" os conjuntos de dados apresentados a ela, determina algumas propriedades dos conjunto de dados e "aprende" a refletir estas propriedades na sua saída. A rede utiliza padrões, regularidades e correlações para agrupar os conjuntos de dados em classes. As propriedades que a rede vai "aprender" sobre os dados pode variar em função do tipo de arquitetura utilizada e da lei de aprendizagem. Por exemplo, Mapa Auto-Organizavél de Kohonen, Redes de Hopfield e Memória Associativa Bidirecional, são algumas métodos de aprendizado não-supervisionado.

Conforme [2] o aprendizado híbrido, por sua vez, consiste de uma combinação dos aprendizados supervisionado e não supervisionado. Um exemplo é o aprendizado por reforço, onde a rede aprende de seu próprio ambiente, a partir dos dados de entrada. A única informação externa que a rede recebe é a indicação de que a resposta fornecida está correta ou não.

APLICAÇÕES

Em geral, as RNAs não apresentam um bom desempenho em tarefas que não são bem executadas por pessoas. Por exemplo, cálculos matemáticos e processamento de transações que exijam rapidez não são adequados para as RNAs e são mais bem executadas pelos computadores convencionais. Entre as áreas de aplicação das RNAs podemos citar[14]:

RECONHECIMENTO DE PADRÕES

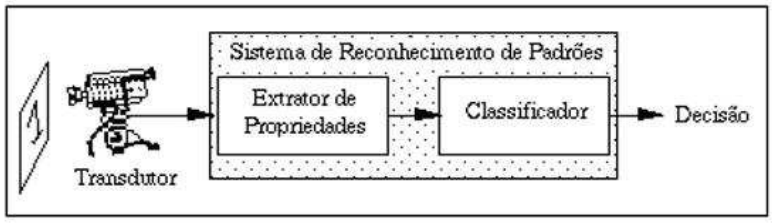
Reconhecimento de padrões é talvez uma das primeiras aplicações de redes neurais. Reconhecimento de padrões é uma tarefa geralmente desempenhada muito melhor usando as capacidades cognitivas do homem do que executando um algoritmo. Por exemplo, seres humanos são excelentes no reconhecimento de rostos, músicas, a caligrafia de

alguém conhecido, etc. Cães são excelentes em reconhecer odores e gatos são capazes de sentir o humor de pessoas, fugindo daquelas que exprimem características agressivas. Isso pode ser atribuído a um sistema bastante desenvolvido de reconhecimento de padrões.[1]

Reconhecimento de Padrões como Classificador

A figura 3 representa esquematicamente um reconhecedor de padrões. O transdutor é munido de um sensor que traduz a forma de energia suporte de informação sobre o objeto (ex: foto-elétrica ou células da retina se informação visual, terminações nervosas do ouvido interno ou microfone se informação sonora) e traduz esta forma de energia para outra capaz de ser processada (neurotransmissores e sinais elétricos de sistema biológico ou elétrico de circuitos artificiais). O processamento inclui geralmente uma primeira fase em que atributos relevantes são selecionados para processamento e este processamento age como uma função, associando ao valor de um conjunto de atributos relevantes um elemento de um conjunto de padrões possíveis, o qual é apresentado como resposta do classificador.[1]

Figura 3: Reconhecimento de padrões



O paradigma mais comum de aprendizado no caso do reconhecimento de padrões é o supervisionado, associado a uma rede direta multi-camadas. Devido a sua disponibilidade, a regra da retro-propagação é frequentemente usada bem como suas variantes. Entretanto bons resultados são obtidos também com o aprendizado competitivo tipo redes de Kohonen. Este último é principalmente interessante quando não se sabe quantas classes possíveis existem a identificar, o que não é o caso do reconhecimento de padrões.[1]

Como exemplos práticos de aplicação podem ser citadas[14]:

- **Data Mining**: Localizar dados em bancos de dados complexos e em sites da Web;
- **Fraudes tributárias**: identificar, localizar e assinalar irregularidades;
- **Serviços financeiros**: identificar padrões de dados sobre o mercado de ações e auxiliar em estratégias de negociação de ações e títulos; escolha e comercialização de commodities, subscrição de hipoteca, precificação de ofertas iniciais ao publico e previsão de taxas de câmbio;
- **Avaliação de pedidos de financiamento**: avaliar a seriedade dos pedidos de financiamento, com base em padrões de informações anteriores (nível de critério do cliente);
- **Previsão de solvência**: avaliando os pontos fortes e fracos de empresas e prevendo possíveis fracassos;
- **Análise de novos produtos**: previsão de vendas e marketing dirigido;
- **Gestão de tarifas aéreas**: procura de assentos e escalas de tripulação;

- **Avaliação de funcionários e candidatos a vagas**: dados dos candidatos comparados às exigências da função de critérios de desempenho;
- **Alocação de recursos baseada em dados históricos e experimentais**: descobrir alocações que maximizem os resultados;
- **Identificação de alvos de aquisição**: prever quais empresas estão mais sujeitas a serem compradas por outras;
- **Validação de assinatura**: comparação e confirmação de assinaturas com amostras no cadastro;
- **Previsões**: antecipar as exigências de comportamento e de pessoal;
- **Detecção de fraudes contra seguradoras**: descobrir padrões de fraude;
- **Detecção de fraudes contra administradoras de cartões de crédito**: análise de padrões de compra para detecção de fraude;

3 MATERIAL E MÉTODOS

3.1 DADOS COLETADOS

Neste trabalho, os métodos multivariados serão aplicados aos dados de hipertensão de um conjunto de idosas na cidade de Curitiba. A coleta foi realizada no período compreendido entre abril e julho de 2006 conforme metodologia descrita em [15] . Com o intuito de realizar uma seleção de amostra aleatória, as seguintes etapas foram realizadas:

1. cadastro das regionais de saúde comunitários existentes no município de Curitiba-PR, obtido mediante parceria com instituições que promovem atividades recreacionais para a população da respectiva regional geográfica;
2. mapeamento de todos os indivíduos cadastrados nas oito regionais do município;
3. alocação aleatória simples dos indivíduos que seriam convidados a participar da pesquisa;
4. visita a regional, explicação dos procedimentos da pesquisa e convite à participação voluntária no estudo.

Depois de realizado o procedimento de seleção amostral, conforme descrito, foi determinado o cronograma para a coleta de dados. A amostra foi constituída de 989 mulheres que estivessem, na data da coleta, com idade cronológica igual ou superior a 60 anos.

Após detalhado esclarecimento sobre os propósitos da investigação, os procedimentos utilizados, os benefícios e os possíveis riscos atrelados, os sujeitos participantes assinaram o termo de consentimento, condicionando sua participação de modo voluntário.

Não foi coletada a variável raça na pesquisa pois não houve entre os pesquisadores o consenso e afirmou-se que poderia ser anti-ético, implicando problemas no julgamento do Comitê de Ética.

O protocolo de pesquisa foi aprovado pelo Comitê de Ética do Setor de Ciências Biológicas da Universidade Federal do Paraná, conforme as normas estabelecidas na Declaração de Helsinque e na Resolução 196/96 do Conselho Nacional de Saúde sobre pesquisa envolvendo seres humanos.

As avaliações foram realizadas no Departamento de Educação Física - Laboratório de Fisiologia do Centro de Pesquisa em Exercício e Esporte, da Universidade Federal do Paraná.

Com o intuito de evitar a influência de variações circadianas todas as avaliações foram realizadas num mesmo período do dia (entre 08:00 e 10:00 horas). Os sujeitos

participantes foram instruídos a não realizar atividade física vigorosa no dia anterior, como também a não ingerir alimento por um período de duas horas antecedendo ao seu início. As avaliações foram realizadas no Departamento de Educação Física, da Universidade Federal do Paraná. Foram coletadas variáveis antropométricas, aplicado o teste de caminhada de seis minutos e também foi administrado o questionário em formato de entrevista.As variáveis antropométicas foram obtidas conforme procedimentos propostos por Lohman et al. (1988).

Para a determinação da estatura (em cm), o indivíduo avaliado permaneceu em posição ortostática com os pés unidos, descalço, utilizando o mínimo possível de roupas. Além disso, deveria manter-se em apnéia inspiratória e com a cabeça orientada em 90^{o} conforme plano de Frankfort, tendo as superfícies do calcanhar, cintura pélvica, cintura escapular e região occipital em contato com o estadiômetro (SANNY, modelo STANDARD, precisão de 0,1 centímetro), o qual encontrava-se fixado a parede.

A massa corporal (MC, em quilogramas) foi mensurada com o individuo avaliado permanecendo em posição ortostática, descalço, e trajando o mínimo possível de roupas. A massa corporal deveria permanecer distribuída entre os membros inferiores durante a permanência na plataforma da balança digital (TOLEDO, modelo 2096 PP; precisão de 0,1 quilogramas). O índice de massa corporal (IMC) foi obtido mediante a utilização do quociente massa corporal/estatura2, onde o valor da massa corporal é expresso em quilogramas e o de estatura em metros.

O teste de caminhada de seis minutos (Tc6) foi conduzido conforme padronização proposta por Rikli e Jones (1999), sendo realizado em uma área retangular de 54,4 metros (18,0 m comprimento x 9,2 m largura). Após breve instrução dos procedimentos do teste, o individuo participante posicionava-se atrás de uma linha que sinalizava o ponto de partida.

Quando o avaliador principal dava o sinal inicial, o sujeito deveria percorrer a maior distância possível dentro de seis minutos, cronometrados (cronômetro marca TIMEX, modelo 85103) por um segundo avaliador.

Foi permitido aos sujeitos reduzir a velocidade durante a realização da caminhada, ou até mesmo finalizar o teste se algum desses sintomas fosse sentido: dispnéia, tontura e dores no peito, cabeça ou pernas. O resultado do teste foi obtido em metros percorridos no tempo de seis minutos. De acordo com Rikli e Jones (1998), o Tc6 apresenta uma considerável correlação com o teste submáximo de esteira ($0{,}71 < r < 0{,}82$), reprodutibilidade teste-reteste ($0{,}88 < r < 0{,}94$) e validade de construto, podendo ser considerado um fidedigno indicador de aptidão cardiorrespiratória em sujeitos idosos.

Para se medir a pressão sangüínea, amarra-se uma faixa tubular de borracha em volta do braço. Liga-se essa faixa a um aparelho que mede a pressão e ela é inflada com ar, enquanto o médico ausculta o pulso arterial na dobra do braço. Aumenta-se a pressão do ar na faixa até que a pulsação não possa mais ser ouvida. Depois disso, esvazia-se a faixa até que o médico comece a ouvir novamente os batimentos do pulso. Nesse momento, a pressão é chamada sistólica. Esvazia-se gradualmente a faixa até que o pulso desapareça

outra vez. Nesse ponto a pressão chama-se diastólica.[23]

Elevações ocasionais da pressão podem ocorrer com exercícios físicos, nervosismo, preocupa ções, drogas, alimentos, fumo, álcool e café.[5]

As mulheres têm, percentualmente, mais hipertensão que os homens, mas eles têm hipertensão mais severa.[6]

Crise hipertensiva é a elevação, repentina, rápida, severa, inapropriada e sintomática da pressão arterial, em pessoa normotensa ou hipertensa. Os órgãos alvo da crise hipertensiva são: os olhos, rins, coração e cérebro. A crise hipertensiva apresenta sinais e sintomas agudos de intensidade severa e grave com possibilidades de deterioração rápida dos órgãos alvo. Pode haver risco de vida potencial e imediato, pois os níveis tensionais estarão muito elevados, superiores a 110 mmHg de pressão arterial diastólica ou mínima.[6]

A hipertensão arterial pode ser sistólica e diastólica (máxima e mínima) ou só sistólica (máxima). A maioria desses indivíduos, 95%, tem hipertensão arterial chamada de essencial ou primária (sem causa) e 5% têm hipertensão arterial secundária a uma causa bem definida.[5]

O aumento de 20mmHg na pressão arterial sistólica ou de 10mmHg na pressão arterial diastólica, em indivíduos entre 40 a 70 anos de idade, dobra o risco para doença cardiovascular. [26]

Outro aspecto que merece consideração é a modificação no perfil da população brasileira com relação aos hábitos alimentares e de vida, que indica uma exposição cada vez mais intensa a riscos cardiovasculares. A mudança nas quantidades de alimentos ingeridos e na própria composição da dieta provocou alterações significativas do peso corporal e distribuição da gordura, com o aumento progressivo da prevalência de sobrepeso ou obesidade da população. Adicione-se a isso a baixa freqüência à prática de atividade física, que também contribui no delineamento desse quadro.[12]

As pesquisas sobre as condições e necessidades dos idosos, no entanto, estão apenas começando, tanto no Brasil como na América Latina.[3] Justifica-se, assim, o presente estudo, que pode contribuir para um melhor conhecimento das representações da doença entre mulheres idosas.

3.2 DESCRIÇÃO DOS DADOS

3.3 Variáveis Coletadas

O banco de dados repassado pela pesquisadora contém 33 variáveis, incluíndo a variável resposta. A descrição das mesmas consta na Tabela 5:

Variável	Nome R
Nascimento	nasc
Sexo	sexo
Bairro	bairro
Grau de Instrução	inst
Tabagismo	tabag
DCV - Doença Cardiovascular	dcv
Data do Teste	dteste
Peso	peso
Estatura	estat
Pressão Arterial Sistólica	pasis
Pressão Arterial Diastólica	padias
Circuferência da Cintura	ccint
Circuferência do Abdomem	cabd
Circuferência do Quadril	cquad
Circuferência da Coxa	ccoxa
Dobra Cutânea Abdominal	dcabd
Dobra Cutânea Supra Ilíaca	dcsupra
Dobra Cutânea Tríceps	dctric
Dobra Cutânea Coxa	dccoxa
FA30 - Resistência muscular de membro superior	fa30
LC30 - Resistência muscular de membro inferior	lc30
TC 6 - Aptidão cardiorespiratória	tc6x
TC6 F - Aptidão cardiorespiratória F	tc6f
Idade	Idade
IMC - Índice de Massa Corporal	imc
RCQ - Razão Cintura Quadril	rcq
Situação Socio Economica	nsecon
Exercicio	exerc
Hipertensao	hipertensao
Diabetes	diabete
Regional-Posto de Saúde	regi
Hipertensão Aferida	hipaf
Soma das Dobras Cutâneas	somadc

Tabela 5: Variáveis Coletadas pela Pesquisadora

3.4 ESTUDO TRANSVERSAL

Antes de iniciarmos as análises, devemos definir qual o tipo de estudo foi realizado, para dessa forma definirmos e aplicarmos métodos estatísticos mais eficientes.

Os estudos observacionais, descritivos e analíticos, podem ser categorizados em transversal (cross-sectional), longitudinal (cohort) e tipo caso-controle:[17]

1. Estudo transversal (estudo de prevalência) fornece uma informação limitada no

tempo - pontual - de uma situação. As medidas ou coletas dos dados são realizadas uma única vez e no mesmo intervalo de tempo;

2. Estudo longitudinal fornece dados acerca de eventos ou mudanças que ocorrem em determinado espaço de tempo. As medidas ou coletas dos dados são realizadas mais de uma vez e em período de tempo diferente. O estudo longitudinal em que grupo de indivíduos é acompanhado por algum tempo é chamado de estudo coorte;

3. Estudo de caso-controle é o tipo de desenho em que se investiga a associação entre a ocorrência de uma doença e a exposição a algum fator suspeito daquela doença. Nesse estudo identifica-se, inicialmente, um grupo de indivíduos com (casos), e sem doença (controle). Depois se investiga no passado causas de diferenças entre as variáveis preditivas que possam explicar por quê os casos adoeceram e os controles não.

Dada às informações acima, concluímos que nosso estudo será transversal.

ESTUDOS TRANSVERSAIS OU CROSS-SECTIONAL

Em estudos transversais coletam-se simultaneamente, de um grupo ou população de indivíduos, informações sobre uma variedade de características que são posteriormente cruzadas em tabelas de contingência. Esta coleta é realizada em um único ponto no tempo e, freqüentemente, o pesquisador não sabe o que ocorreu antes desse ponto. A obtenção da prevalência da doença, ou seja, da proporção do grupo com a doença no momento em que foi realizada a coleta, é um dos objetivos desses estudos. Constitui outro interesse, em geral, a investigação de potenciais relações causais entre os fatores suspeitos serem de risco e a doença.[10]

Os estudos transversais podem ser vistos como avaliações fotográficas de grupos ou populações de indivíduos. O termo transversal é usado para indicar que os indivíduos estão sendo estudados em um ponto no tempo (corte transversal). O interesse está em avaliar a associação entre as respostas obtidas. Nesses estudos é comum considerar algumas das variáveis como fatores.[10]

3.4.1 Prevalência

Nos estudos transversais, a avaliação não é feita ao longo do tempo, mas somente em um único ponto (momento) no tempo. Alguns dos indivíduos neste ponto do tempo apresentarão a resposta e outros não. Não é observado, portanto, casos novos ao longo do tempo, mas somente os casos existentes naquele momento específico. A medida adequada é, desse modo, a prevalência, isto é, a proporção de indivíduos do grupo com resposta positiva naquele momento específico do tempo, ou seja:[10]

Para os dados de hipertensão temos 504 mulheres com a presença hipertensão e 485 com ausência. Então a prevalência de hipertensão na amostra é:

Prevalência = $\frac{504}{989}$ = 0,5096056 ou em percentual **50,96%**

3.5 AMOSTRAGEM

Antes de tudo, é preciso garantir que a amostra ou amostras que serão usadas sejam obtidas por processos adequados. Se erros palmares forem cometidos no momento de selecionar os elementos da amostra, o trabalho todo ficará comprometido e os resultados finais serão provavelmente bastante incorretos. Devemos, portanto, tomar especial cuidado quanto aos critérios que serão usados na seleção da amostra.[4]

Na realização de qualquer estudo quase nunca é possível examinar todos os elementos da população de interesse. A inferência estatística nos dá elementos para generalizar, de maneira segura, as conclusões obtidas da amostra para a população.[29]

O que é necessário garantir, em suma, é que a amostra seja *representativa* da população. Isso significa que, a menos de certas pequenas discrepâncias inerentes à aleatoriedade sempre presente, em maior ou menor grau, no processo de amostragem, a amostra deve possuir as mesmas características básicas da população, no que diz respeito às variáveis que desejamos pesquisar.[4]

É errôneo pensar que, caso tivéssemos acesso a todos os elementos da população, seríamos mais precisos. Os erros de coleta e manuseio de um grande número de dados são maiores do que as imprecisões a que estamos sujeitos quando generalizamos, via inferência, as conclusões de uma amostra bem selecionada. A textitpopulação-alvo é a população sobre a vamos fazer inferências baseadas na amostra. Uma causa freqüente de levantamentos ruins é a falta de cuidado com que a população-alvo é definida. Uma das formas de se conseguir representatividade é fazer com que o processo de escolha da amostra seja, de alguma forma, aleatório. Além disso, a aleatoriedade permite o cálculo de estimativas dos erros envolvidos no processo de inferência.[29]

Nossa pesquisadora, apesar de haver coletado aleatoriamente os dados, não fez um cálculo amostral inicial. Dessa forma devemos verificar a representatividade da amostra apresentada por ela. Escolhemos o método da proporção.

Proporção

Muitas vezes, especialmente em medicina, a variável de interesse não é qualitativa, mas sim dicotômica, isto é, uma variável que só assume dois valores possíveis. Este é o caso quando estamos interessados em saber se uma característica está presente ou não.[28]

No nosso caso trata-se da hipertensão: hipertenso ou não.

Dessa forma, segundo SOARES 1991, se a pergunta é do tipo SIM-NÃO e queremos, com nível (1 - α) de certeza, que a proporção estimada esteja, no máximo, a uma distância d da proporção verdadeira, ou seja, se queremos que

$$Pr[|p - \hat{p}|] \leq d = 1 - \alpha \tag{46}$$

o valor de n é dado por

$$n = \frac{z^2 PQ}{d^2} \tag{47}$$

onde

n = dimensão que pretendemos saber

Z = limiar de confiança que pretendemos ($z = Z_1 - \frac{\alpha}{2}$)

P = proporção conhecida ou 50% quando não a conhecemos ($Q = 1 - P$)

d = maior desvio aceitável, ou seja, erro que estamos dispostos a aceitar

Importância de algumas variáveis coletadas

Nesta seção vamos indicar conforme [7] a importância de algumas variáveis para o problema da hipertensão. Os fatos enumerados a seguir devem ser levados em consideração na formulação de um modelo probabiliístico.

1. Peso corpóreo - O excesso de peso aumenta de duas a seis vezes o risco de hipertensão, enquanto a diminuição de peso em normotensos reduz a pressão e a incidência de hipertensão. IMC entre 20 e 25 é considerado peso normal, acima de 25 o indivíduo é considerado obeso.

2. Atividade física - Há relação inversa entre o grau de atividade física e a incidência de hipertensão. O exercício físico regular reduz a pressão sistólica/diastólica em 3/2 mmHg em normotensos, sendo a queda proporcional à pressão arterial inicial, o que recomenda sua prática.

3. Dislipidemias - Hipercolesterolemia e hipertrigliceridemia com HDL-colesterol baixo são importantes fatores de risco cardiovascular. As metas atuais do controle do perfil lipídico são classificadas de acordo com o risco de um evento coronário agudo em dez anos.

4. Intolerância à glicose e diabete melito - São condições freqüentemente associadas à hipertensão arterial, favorecendo a ocorrência de doenças cardiovasculares e complicações do diabete.

5. Tabagismo - Recomenda-se a interrupção do tabagismo porque ele se associa a maior incidência e mortalidade das doenças coronária, cerebrovascular e vascular de extremidades.

4 RESULTADOS E DISCUSSÕES

Antes da apresentação dos resultados, algumas considerações no tratamento dos dados serão feitas conforme a lista que segue.

1. Nascimento, será analisado pela idade;
2. Sexo, a amostra só conta com mulheres;
3. Bairro, será analisado pela regional;
4. Data do Teste, esse variável foi colocada apenas para o controle da pesquisadora, que entende que ela não seja de importância para a análise;
5. Peso e Estatura, serão analisados pelo IMC (índice de massa corporal);
6. Pressão Arterial Sistólica, Diastólica e Hipertensão Aferida, essas variáveis estão ligadas diretamente com a variável resposta;
7. Circunferência da Cintura e do Quadril, serão analisados pelo RCQ (razão cintura quadril);
8. Soma das Dobras Cutâneas, analisaremos as dobras cutâneas individualmente, uma vez que foi disponibilizada a informação.

Alguns dos métodos de classificação foram avaliados em relação ao desempenho quando aplicados a um conjunto de dados que não foi utilizado para o ajuste do modelo. Para cumprir esta finalidade, os dados foram particionados em duas partes: Treinamento (90%) e Teste (10%). A capacidade preditiva dos métodos é avaliada única e exclusivamente no conjunto de teste.

4.1 AMOSTRAGEM

Em nosso caso já temos n, dessa forma isolaremos o maior desvio aceitável, ou seja, o erro máximo que estamos dispostos a aceitar. Estabelecemos que no máximo ele poderá atingir será $0,05$ para ser representativo.

Caso 1 - Confiança de 95% - $\alpha = 5\%$ - $P = 50\%$

$$989 = \frac{(1,96)^2(0,5)(0,5)}{d^2}$$

$$d^2 = \frac{0,9604}{989}$$
$$d = \sqrt{0,0009711}$$
$$d = 0,0311622$$

Caso 2 - Confiança de 99% - $\alpha = 1\%$ - $P = 50\%$

$$989 = \frac{(2,58)^2(0,5)(0,5)}{d^2}$$
$$d^2 = \frac{1,6641}{989}$$
$$d = \sqrt{0,0016826}$$
$$d = 0,0410196$$

Dado os cálculos acima, uma amostra de 989 mulheres com mais de 60 anos apresenta um erro menor que 5% com um nível de confiança de 99%, sendo assim, podemos dizer que a amostra é representativa a população.

4.2 ANÁLISE DESCRITIVA

Realizamos a análise descritiva nos dados por meio de medidas-resumo, obtendo média, mediana, mínimo, máximo, variância e desvio padrão, além de realizar análise gráficas individuais. Durante essa análise verificamos se nenhuma das variáveis apresentavam valores inconsistentes, como por exemplo um valor zero no IMC, isso poderia interferir nos resultados dos modelos de classificação. Felizmente, não encontramos observações suspeitas. Abaixo, apresentamos alguns resultados, obtidos através da utilização do software R ([21]).

	Min	1o Q	Mediana	Média	3o Q	Max	Var	DP
peso	39.00	60.00	67.40	68.30	75.29	132.10	142.55	11.94
estat	132.0	150.5	155.0	154.7	159.0	180.00	38.117	6.174
pasis	100.0	120.0	130.0	132.6	140.0	190.00	199.61	14.13
padias	50.00	75.00	80.00	80.63	90.00	120.00	85.149	9.227
ccint	35.00	80.00	86.00	87.03	94.00	135.00	112.26	10.59
cabd	56.00	91.00	98.00	98.39	105.0	150.00	118.14	10.86
cquad	73.00	95.00	100.0	100.8	106.0	141.00	89.141	9.441
ccoxa	34.00	43.00	47.00	47.38	51.00	98.000	35.703	5.975
dcabd	5.000	32.00	40.00	39.91	48.00	90.000	123.89	11.13
dcsupra	5.000	28.00	35.00	34.65	42.00	63.000	114.12	10.68
dctric	6.000	20.00	25.00	25.68	30.00	56.000	64.516	8.032
dccoxa	4.000	24.00	30.00	30.00	36.00	62.000	99.528	9.976
fa30	5.000	12.00	14.00	14.27	16.00	36.000	11.626	3.409
lc30	0.000	11.00	13.00	12.87	15.00	22.000	7.9138	2.813
tc6x	106.3	449.0	499.2	492.3	544.3	717.10	6170.5	78.55
idade	60.00	64.40	68.80	69.37	73.50	87.900	36.245	6.020
imc	17.00	25.42	28.00	28.52	31.20	51.600	22.231	4.715
rcq	0.330	0.820	0.860	0.863	0.910	1.0900	0.0051	0.072
exerc	0.000	0.738	2.047	2.474	3.146	15.930	5.0848	2.255
somadc	35.00	128.0	156.0	154.6	179.0	275.00	1350.1	36.74

Tabela 6: Análise Descritiva dos Dados - Variáveis Quantitativas

Min- Mínimo /1oQ - Primeiro Quartil / 3oQ - Terceiro Quartil / Max - Máximo /Var - Variância/ DP - Desvio Padrão

4.3 ANÁLISE DE CLUSTER

Com auxílio do software R foi feita a clusterização pelo método das k-médias.

A proposta inicial foi variar os tamanhos dos clusters para se encontrar grupos com uma quantidade de elementos que não fosse muito pequena (ex: abaixo de 20). Estes critérios, embora arbitrários, procuram dar representatividade ao cluster no momento em que for feito algum tipo de comparação.

```
> clu8 = kmeans(dadosclus, 8)
> clu8$size

[1]  57  55 150 157 128  13 190 140

> clu7 = kmeans(dadosclus, 7)
> clu7$size

[1] 202 140 222  64 192  13  57
```

```
> clu6 = kmeans(dadosclus, 6)
> clu6$size

[1] 232 129  33  73 234 189

> clu5 = kmeans(dadosclus, 5)
> clu5$size

[1] 180 291 278  33 108

> clu4 = kmeans(dadosclus, 4)
> clu4$size

[1] 353 218 261  58

> clu3 = kmeans(dadosclus, 3)
> clu3$size

[1] 347 415 128
```

Em função dos resultados acima expostos, vamos determinar $k = 4$ como um número satisfatório de clusters. Para esta separação, apresentamos os centróides de cada cluster, ou seja, o vetor com as médias das variáveis.

Cluster	pasis	padias	cabd	ccoxa	dcabd	dcsupra	dctric	dccoxa
1	132.3258	81.30312	98.94618	47.78754	41.27195	35.97167	26.32861	30.0311
2	130.9633	79.93119	95.50917	47.32110	38.34862	33.99083	24.66972	29.5550
3	133.6322	80.82375	99.11877	46.90805	39.22605	33.70115	25.40613	30.0191
4	135.2069	78.27586	102.51724	47.17241	40.58621	33.34483	26.70690	31.3793

Cluster	fa30	lc30	tc6x	Idade	imc	rcq	exerc
1	14.78470	13.226629	509.2652	68.53768	28.97167	0.8615014	2.491587
2	15.54128	14.068807	582.7913	66.39450	26.96606	0.8444037	3.380495
3	13.00383	12.053640	433.8126	71.58238	28.75479	0.8792720	1.973988
4	11.98276	9.896552	311.6983	75.67586	30.54655	0.8765517	1.204121

O *cluster 4* destaca-se por apresentar a maior pressão sistólica média, combinada com a maior circunferência média de abdomem, maior dobra cutânea de tríceps e coxa, maior idade e maior IMC. Além disso, apresenta menor resistência musclular dos membros superiores e inferiores, menor aptidão cardiorespiratória e é o que possui menos intensidade de prática de exercícios.

O *cluster 2* apresenta a menor pressão sistólica média e a segunda menor pressão diastólica média. É o grupo com menor circunferência média de abdomem, menor dobra cutânea do tríceps, coxa e abdomem. Esse *cluster* tem a menor idade média, menor IMC, menor razão cintura quadril combinados com a maior intensidade de prática de exercícios, maior aptidão cardiorespiratória e maior resistência muscular nos membros superiores e inferiores.

4.4 ÁRVORES DE DECISÃO

Para gerar as árvores de classificação utilizou-se o pacote tree do software estatístico R. Pelo uso da árvore de classificação é possível averiguar variáveis importantes na discriminação de hipertensos e não hipertensos.

O algoritmo utilizado para a classificação foi o CART com parâmetros padrões do pacote R e a saída computacional é ilustrada em 4.

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 890 1234.0 0 ( 0.5000 0.5000 )
  2) cabd < 93.5 298  392.5 0 ( 0.6309 0.3691 )
    4) tc6x < 492.7 123  170.5 0 ( 0.5041 0.4959 ) *
    5) tc6x > 492.7 175  207.5 0 ( 0.7200 0.2800 ) *
  3) cabd > 93.5 592  810.4 1 ( 0.4341 0.5659 )
    6) dctric < 34.5 481  665.1 1 ( 0.4699 0.5301 ) *
    7) dctric > 34.5 111  131.5 1 ( 0.2793 0.7207 ) *
```

Figura 4: Saída Computacional do Pacote R para Árvore de Classificação Construída

O Resultado obtido pelo uso da árvore de classificação gerou um conjunto de 4 regras de classificação:

1. SE Cintura do Abdomen menor que 93,5 E Aptidão Cardio Respiratória menor que 492,7 ENTÃO: PROBABILIDADE DE HIPERTENSÃO É 49,9 %
2. SE Cintura do Abdomen meno que 93,5 E Aptidão Cardio Respiratória maior que 492,7 ENTÃO: PROBABILIDADE DE HIPERTENSÃO É 28 %
3. SE Cintura do Abdomen maior que 93,5 E Dobra Cutânea do Triceps menor que 34,5 ENTÃO:PROBABILIDADE DE HIPERTENSÃO É 53,01 %
4. SE Cintura do Abdomen maior que 93,5 E Aptidão Cardio Respiratória maior que 34,5 que ENTÃO:PROBABILIDADE DE HIPERTENSÃO É 72,07 %

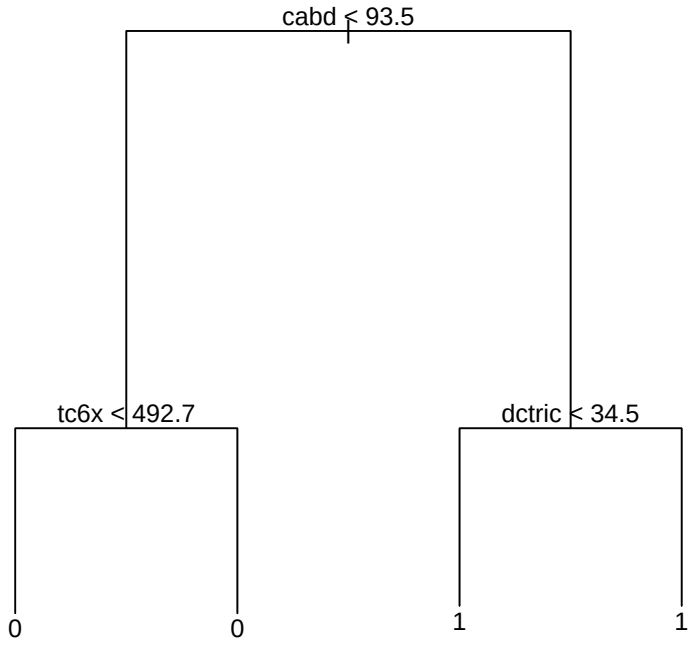


Figura 5: Árvore de Classificação para idosas hipertensas

Os resultados apresentados pela árvore de classificação vão de encontro a literatura específica sobre esta doença que relata aumento na probabilidade de hipertensão em pessoas acima do peso e devemos ressaltar que as idosas com maior aptidão cardiorespiratória, provavelmente por realizarem algum tipo de atividade física, apresentam menor probabilidade de manifestação da hipertensão.

	PREDITOS	
OBSERVADO	hipertenso	não hipertenso
hipertenso	40	22
não hipertenso	19	18

Tabela 7: Frequência Absoluta das Classificações no Conjunto de Teste

	PREDITOS	
OBSERVADO	hipertenso	não hipertenso
hipertenso	67,8%	32,2%
não hipertenso	55%	45%

Tabela 8: Frequência Relativa das Classificações no Conjunto de Teste

4.5 REGRESSÃO LOGÍSTICA

Utilizamos o *software R* para a realização do ajuste da regressão logística com estimação feita pelo método da máxima verossimilhança.

Inicialmente, foi ajustado o seguinte modelo de regressão logística para o conjunto de dados selecionados, o qual chamou-se de principal:

$$\begin{aligned} logit(\theta(x)) \quad = \quad & beta_0 + beta_1 * inst + beta_2 * tabag + \\ & beta_3 * dcv + beta_4 * cabd + beta_5 * ccoxa + \\ & beta_6 * dcabd + beta_7 * dcsupra + beta_8 * dctric + \\ & beta_9 * dccoxa + beta_{10} * fa30 + beta_{11} * lc30 + \\ & beta_{12} * tc6x + beta_{13} * tc6f + beta_{14} * idade + \\ & beta_{15} * imc + beta_{16} * rcq + beta_{17} * nsecon + \\ & beta_{18} * exerc + beta_{19} * diabete \end{aligned}$$

A partir do modelo principal, foram utilizados os critérios de escolhas anteriormente descritos que resultou num subconjunto de modelos que apresentou melhores resultados nestes critérios.

Modelo II

$$\begin{aligned} ModeloII(\theta(x)) &= beta_0 + beta_1 * inst + beta_2 * dcv + \\ & beta_3 * cabd + beta_4 * dctric + beta_5 * lc30 + \\ & beta_6 * rcq + beta_7 * diabete \end{aligned}$$

Modelo IV

$$\begin{aligned} ModeloIV(\theta(x)) &= beta_0 + beta_1 * inst + beta_2 * dcv + \\ & beta_3 * dctric + beta_4 * lc30 + beta_5 * imc + \\ & beta_6 * rcq + beta_7 * diabete \end{aligned}$$

Modelo V

$$\begin{aligned} ModeloV(\theta(x)) &= beta_0 + beta_1 * inst + beta_2 * dcv + \\ & beta_3 * lc30 + beta_4 * imc + beta_5 * rcq + \\ & beta_6 * diabete \end{aligned}$$

Modelo VI

$$\begin{aligned} ModeloVI(\theta(x)) &= beta_0 + beta_1 * dcv + beta_2 * lc30 + \\ & beta_3 * imc + beta_4 * rcq + beta_5 * diabete \end{aligned}$$

Modelo	RD	GL	AIC	QL	p-valor	QP	p-valor	OR	LI	LS
Modelo II	1149,2	882	1165,2	1149,20	2,78000E-09	88,90	0,429	2,3478	1,03	5,34
Modelo IV	1150,7	882	1166,7	1150,70	2,32000E-09	889,30	0,425	2,1153	0,93	4,81
Modelo V	1157,2	883	1171,2	1157,20	1,22000E-09	889,5	0,433	2,6361	1,14	6,09
Modelo VI	1166,4	884	1178,4	1166,40	4,51000E-10	890,50	0,432	2,9615	1,27	6,92

Tabela 9: Comparação entre os Modelos – Parte I

Modelo	Teste	Hipert=1	Hipert=0	Totais	V Pred	V Pred +	V Pred -	Falsos +	Falsos -	Sensib	Especif
Mod II	Hipert=1	36	23	59	0,606	0,692	0,511	0,400	0,390	0,610	0,600
	Hipert=0	16	24	40							
	Totais	52	47	99							
Mod IV	Hipert=1	33	26	59	0,586	0,688	0,490	0,375	0,441	0,559	0,625
	Hipert=0	15	25	40							
	Totais	48	51	99							
Mod V	Hipert=1	33	26	59	0,606	0,717	0,509	0,325	0,441	0,559	0,675
	Hipert=0	13	27	40							
	Totais	46	53	99							
Mod VI	Hipert=1	33	26	59	0,616	0,733	0,519	0,300	0,441	0,559	0,700
	Hipert=0	12	28	40							
	Totais	45	54	99							

Tabela 10: Comparação entre os Modelos – Parte II

Analisando os resultados apresentados escolhemos o Modelo VI, para representar a regressão logística nessa análise.

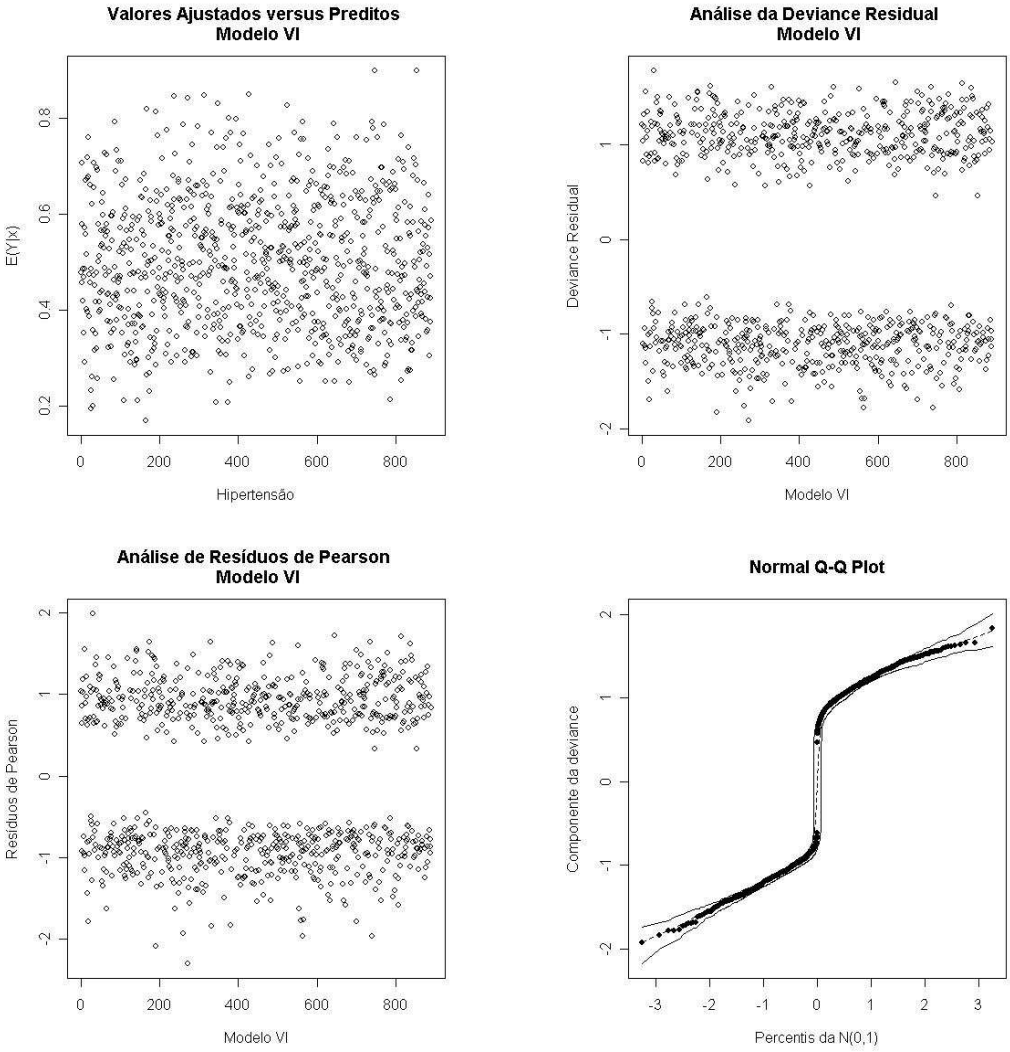


Figura 6: Diagnóstico dos Resíduos

Coeficientes	Estimativas	Erro Padrão
Intercepto	-3.57713	1.02838
dcv	0.32542	0.14090
lc30	-0.07214	0.02528
imc	0.07167	0.01597
rcq	2.59723	1.03857
diabete	0.48291	0.19658

Tabela 11: Parâmetros Estimados – Modelo VI

Na sequência apresentamos os gráficos para o diagnóstico dos resíduos obtidos após o ajuste do modelo VI, considerado o melhor dentre os ajustados.

Conforme a Figura 6, não há evidência de padrão no comportamento dos resíduos e, portanto, assumimos que o modelo IV apresenta resultados satisfatórios.

4.6 ANÁLISE DISCRIMINANTE

Utilizamos o *software R* para a realização dos cálculos da análise linear discriminante. Na Tabela 12 podem-se observar os coeficientes da função linear discriminante.

Variáveis	Coeficientes da Função Linear Discriminante (LD1)
cabd	0.027639397
ccoxa	-0.042793393
dcabd	0.010266394
dcsupra	-0.039468136
dctric	0.061384790
dccoxa	-0.010297142
fa30	0.008274048
lc30	-0.106611417
tc6x	-0.002487503
Idade	-0.011472827
imc	0.098871140
rcq	4.745425728
exerc	0.082519760

Tabela 12: Coeficientes da Análise Linear Discriminante

A partir do modelo estimado e apresentado na Tabela 12 verificamos no conjunto de teste o comportamento, em termos preditivos, deste modelo.

TESTE	Hipert=1	Hipert=0	Totais
Hipert=1	34	25	59
Hipert=0	16	24	40
Totais	50	49	99

Tabela 13: Resultados através da utilização do Modelo ALD

Valor Pred	Valor Pred +	Valor Pred -	Falsos +	Falsos -	Sensibilidade	Especificidade
0,586	0,68	0,49	0,4	0,424	0,5763	0,6

Tabela 14: Resultados através da utilização do Modelo ALD

4.7 REDES NEURAIS ARTIFICIAIS

Uma das principais vantagens das RNA é o poder preditivo para problemas em que há alta complexidade na relação entre as variáveis. Para o ajuste da Rede foram utilizados os dados padronizados e a topologia escolhida foi 13-20-1 para uma rede do tipo Feed-Forward.

De fato esta caracteriza-se como uma rede altamente complexa e repleta de não-linearidades. A qualidade da classificação da RNA é apresentada nas Tabelas 15 e 16 que ilustram uma alta taxa de acertos no diagnóstico dos hipertensos.

	PREDITOS	
OBSERVADO	hipertenso	não hipertenso
hipertenso	43	16
não hipertenso	15	25

Tabela 15: Frequência Absoluta das Classificações no Conjunto de Teste

	PREDITOS	
OBSERVADO	hipertenso	não hipertenso
hipertenso	72,88%	27,12%
não hipertenso	37,50%	62,50%

Tabela 16: Frequência Relativa das Classificações no Conjunto de Teste

5 CONCLUSÕES

Durante o desenvolvimento deste trabalho validou-se a representatividade da amostra em estudo e concluiu-se que a amostra é representativa. Sendo assim, podemos afirmar que a amostra representa a população de mulheres curitibanas com mais de 60 anos, visto que a amostra de 989 mulheres com essas características apresenta um erro menor que 5

Observou-se também que os dados se encaixam no estudo transversal, pois fornecem uma informação limitada no tempo - pontual - de uma situação. As medidas ou coletas dos dados foram realizadas uma única vez e no mesmo intervalo de tempo.

Diante dos resultados obtidos na análise de cluster verifica-se que o grupo de idosas com maior pressão sistólica média também apresenta as maiores medidas de circunferência de abdomem, dobra cutânea de tríceps e coxa, é o grupo de maior idade e maior IMC. Alem disso, apresenta menor resistência muscular tanto nos membros superiores quanto inferiores, e apresenta menor aptidão cardiorespiratória, associada a baixa de prática de exercícios. Em contraste a esse grupo, forma-se outro com as pessoas com a menor pressão sistólica média, que é formado pelas mais jovens, com menor IMC, menor circunferência média de abdomem, menor dobra cutânea do tríceps, coxa e abdomem e menor razão cintura quadril. Esse cluster também apresenta a maior intensidade de pratica de exercícios associada a maior aptidão cardiorespiratória e maior resistência muscular nos membros superiores e inferiores. Em virtude disso, é possível observar que a hipertensão está associada ao aumento da idade, a obesidade e a falta de prática de exercícios físicos.

Analisando a árvore de classificação confirma-se o observado na análise de cluster. Fica evidente o aumento da probabilidade de hipertensão em pessoas acima do peso e observa-se que as idosas com maior aptidão cardiorespiratória, provavelmente por realizarem algum tipo de atividade física, apresentam menor probabilidade de manifestação da hipertensão.

Na análise realizada por regressão logística observou-se como fator de risco a razão cintura quadril e o IMC, que estão associados à obesidade. Além disso, destacam-se como fatores de risco para hipertensão a presença de doenças cardiovasculares e diabetes. Por outro lado, o aumento de resistência muscular nos membros superiores esta associado a redução da probabilidade de ocorrência de hipertensão.

Na análise discriminante também aparecem como variáveis importantes para discriminação as que estão associadas à obesidade, como IMC, razão cintura quadril, circunferências do abdomem e coxa, dobras cutâneas do tríceps, abdomem, coxa e supra ilíaca. Outro fator discriminante é a idade. E em concordância com os outros métodos já mencionados e com a literatura específica da área aparecem fatores como a intensidade de prática de exercícios, aptidão cardiorespiratória e resistência dos menbros superiores e inferiores.

Comparando os modelos de previsão ajustados por Regressão Logística, Árvores de Classificação, Análise Discriminante Linear de Fisher e Redes Neurais Artificiais,

observa-se que o modelo com maior sensibilidade (Classifica como hipertensa quando a idosa realmente é hipertensa) é o obtido por Redes Neurais Artificiais, com sensibilidade de 72,88Observada a sensibilidade e especificidade de cada modelo podemos indicar como "melhor"modelo preditivo o ajustado por Redes Neurais Artificias, porém esse modelo é bastante complexo dada sua topologia (13-20-1) e apesar de sua alta taxa de "acertos" seus parâmetros não possuem interpretações práticas. Modelos como os ajustados por Árvores de Classificação, Análise Discriminante e Regressão logística são mais simples e de mais fácil interpretação.

Referências

[1] Jorge M. Barreto. Introdução às Redes Neurais Artificiais. 04 2002.

[2] Nair Cristina Margarido Brondino. Estudo da Influência da Acessibilidade no Valor de Lotes Urbanos Através do Uso de Redes Neurais. 1999.

[3] Fernanda Carvalho, Rodolpho Telarolli Junior, and José Cândido Monteiro da Silva Machado. Uma investigação antropológica na terceira idade: concepções sobre a hipertensão arterial. *Cadernos de Saúde Pública*, 14:617 – 621, 07 1998.

[4] Pedro Luís de Oliveira Costa Neto. *Estatística*. Editora Edgard Blücher, São Paulo, 1991. 14ª reimpressão.

[5] ABC da Saúde. Hipertensão Arterial - Introdução. 11 2007. http://www.abcdasaude.com.br/artigo.php?244.

[6] ABC da Saúde. Hipertensão Arterial - Investigação Clínica e Laboratorial. 11 2007. http://www.abcdasaude.com.br/artigo.php?245.

[7] Sociedade Brasileira de Nefrologia. SBN - Capítulo 9 - Prevenção da Hipertensão e dos Fatores de Risco Associados. 11 2007. http://www.sbn.org.br/Diretrizes/HA/Capitulo%2009%20diretrizes%20corrigido.pdf.

[8] Lúcia Pereira Barroso e Rinaldo Artes. *Minicurso Análise Multivariada*. 2003.

[9] Medicina e Saúde. Pressão Arterial. 11 2007.

[10] Suely Ruiz Giolo. *Análise de Dados Categóricos*. Curitiba, 2006. Apostila e Notas de Aula.

[11] Rodolfo Hoffmann and Sônia Vieira. *Análise de Regressão: uma introdução a econometria*. Hucitec - USP, São Paulo, 1977.

[12] Paulo César B. Veiga Jardim, Maria do Rosário Peixoto Gondim, Estelamaris Tronco Monego, Humberto Graner Moreira, Priscila Valverde de Oliveira Vitorino, Weimar Kunz Sebba Barroso Souza, and Luiz César Nazário Scala. Hipertensão arterial e alguns fatores de risco em uma capital brasileira. *Arquivos Brasileiros de Cardiologia*, 88:452 – 457, 04 2007.

[13] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 1998. Sixth Edition.

[14] Daniel Karrer, Renato Florido Cameira, André Strauss Vasques, and Marcos de Almeida Benzecry. Redes Neurais Artificiais: Conceitos e Aplicações.

[15] Maressa Priscila Krause, Tatiane Hallage, Cristiane Petra Miculis, Elisa Cesar Ribeiro dos Santos, Cosme Franklin Buzzachera, and Sergio Gregorio da Silva.

[16] Cecília Amaro de Lolio, Júlio César Rodrigues Pereira, Paulo Andrade Lotufo, and José Maria Pacheco de Souza. Hipertensão arterial e possíveis fatores de risco. *Revista de Saúde Pública*, 27:357 – 362, 10 1993.

[17] Bráulio Luna Filho. Seqüência básica na elaboração de protocolos de pesquisa. *Arquivos Brasileiros de Cardiologia*, 71:735 – 740, 12 1998.

[18] José Simeão de Medeiros. Bancos de Dados Geográficos e Redes Neurais Artificiais: Tecnologias de Apoio a Gestão do Território. 07 1999.

[19] Gilberto A Paula. *Modelos de Regressão com Apoio Computacional*. Universidade de São Paulo, São Paulo, 2004. www.ime.usp.br/ giapaula.

[20] Janete Pessuto and Emília Campos de Carvalho. Fatores de risco em indivíduos com hipertensão arterial. *Revista Latino-Americana de Enfermagem*, 6:33 – 39, 01 1998.

[21] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[22] Reliasoft. Relia Soft Brasil - Conceitos de Confiabilidade: Estimador de Máxima Verossimilhança (MLE). 11 2007. http://www.reliasoft.com.br/hotwire/edicao2/conceito2.htm.

[23] Robert E Rothenberg. *Enciclopédia Médica ilustrada para uso no lar - Vol 1*. Editora Abril, São Paulo, 1976. Tradução: Camargo, Marisis Aranha.

[24] Subhash Sharma. *Applied Multivariate Techniques*. 1996.

[25] Jorge Luis Lima Silva and Solange Lourdes de Souza. Fatores de risco para hipertensão arterial sistêmica versus estilo de vida docente. *Revista Eletrônica de Enfermagem* , 06:330 – 335, 2004.

[26] Manuel Simão. Hipertensão arterial e fatores de risco associados: estudo entre universitários da cidade de Lubango-Angola. Master's thesis, USP, 2005. http://www.teses.usp.br/teses/disponiveis/22/22132/tde-13092005-105607.

[27] Neural Site. Redes Neurais. 11 2007. http://www.din.uem.br/ia/neurais/.

[28] José Francisco Soares and Flávio Celso Bartan. *Métodos Estatísticos em Medicina e Biologia*.

[29] José Francisco Soares, Alfredo Alves de Farias, and Cibele Comini Cesar. *Introdução a Estatística*. Editora Guanabara Koogan, Rio de Janeiro, 1991.

[30] Jonny Arruda Souza, Luis Felipe Snell Zanettini, Marco Tulio Zanettini, Rodrigo Boldo, and Renan Stoll Moraes. Prevalência de hipertensão arterial e fatores de risco associados em trabalhadores de uma instituição de ensino superior. *Revista da AMRIGS*, 49:226 – 232, 2005.

[31] Maria Paula do Amaral Zaitune, Marilisa Berti de Azevedo Barros, Chester Luiz Galvão César, Luana Carandina, and Moisés Goldbaum. Hipertensão arterial em idosos: prevalência, fatores associados e práticas de controle no Município de Campinas, São Paulo, Brasil. *Cadernos de Saúde Pública*, 22:285 – 294, 02 2006.

ANEXOS

COMANDOS R PARA ANÁLISE LINEAR DISCRIMINANTE

```
> mADF <- lda(factor(hipertensao) ~ ., ADF)
> mADF


Call:
lda(factor(hipertensao) ~ ., data = ADF)

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
       cabd    ccoxa    dcabd  dcsupra   dctric   dccoxa     fa30     lc30
0  96.18652 46.99101 38.42247 33.58202 24.31910 29.25393 14.40899 13.22247
1 100.58876 47.75955 41.40000 35.71685 27.03371 30.74382 14.12135 12.52135
      tc6x    Idade      imc       rcq    exerc
0 501.8915 69.14202 27.57303 0.8530787 2.468266
1 482.6539 69.59955 29.46584 0.8739326 2.478989

Coefficients of linear discriminants:
                LD1
cabd     0.027639397
ccoxa   -0.042793393
dcabd    0.010266394
dcsupra -0.039468136
dctric   0.061384790
dccoxa  -0.010297142
fa30     0.008274048
lc30    -0.106611417
tc6x    -0.002487503
Idade   -0.011472827
imc      0.098871140
rcq      4.745425728
exerc    0.082519760
```

Aplicação do modelo na base de testes

```
> preditosteste <- predict(mADF, newdata = dadostest, type = "class")
> mADFYT <- dadostest$hipertensao
> mADFAT <- preditosteste$class
> MADFT <- data.frame(mADFYT, mADFAT)
```

```
> write.table(MADFT, "MADFTEST")
> table(mADFYT, mADFAT)

      mADFAT
mADFYT  0  1
     0 24 16
     1 25 34
```

COMANDOS R PARA ÁRVORES DE DECISÃO

```
> require(tree)

> arv.hipertensao = tree(factor(hipertensao) ~ inst + tabag + dcv +
+     cabd + ccoxa + dcabd + dcsupra + dctric + dccoxa + fa30 +
+     lc30 + tc6x + tc6f + Idade + imc + rcq + nsecon + exerc +
+     diabete + regi + somadc, data = dadostrein)
> arv.hipertensao

node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 890 1234.0 0 ( 0.5000 0.5000 )
  2) cabd < 93.5 298  392.5 0 ( 0.6309 0.3691 )
    4) tc6x < 492.7 123  170.5 0 ( 0.5041 0.4959 ) *
    5) tc6x > 492.7 175  207.5 0 ( 0.7200 0.2800 ) *
  3) cabd > 93.5 592  810.4 1 ( 0.4341 0.5659 )
    6) dctric < 34.5 481  665.1 1 ( 0.4699 0.5301 ) *
    7) dctric > 34.5 111  131.5 1 ( 0.2793 0.7207 ) *

> plot(arv.hipertensao)
> text(arv.hipertensao)

> arv.hipaf = tree(factor(hipaf) ~ inst + tabag + dcv + cabd +
+     ccoxa + dcabd + dcsupra + dctric + dccoxa + fa30 + lc30 +
+     tc6x + tc6f + Idade + imc + rcq + nsecon + exerc + diabete +
+     regi + somadc, data = dadostrein)
> arv.hipaf

node), split, n, deviance, yval, (yprob)
      * denotes terminal node
```

```
 1) root 890 1223.00 1 ( 0.44607 0.55393 )
   2) cabd < 93.5 298  406.60 0 ( 0.57383 0.42617 )
     4) tc6x < 457.5 77  103.00 1 ( 0.38961 0.61039 )
       8) exerc < 2.8645 55   76.23 0 ( 0.50909 0.49091 ) *
       9) exerc > 2.8645 22   13.40 1 ( 0.09091 0.90909 ) *
     5) tc6x > 457.5 221  289.30 0 ( 0.63801 0.36199 ) *
   3) cabd > 93.5 592  787.30 1 ( 0.38176 0.61824 )
     6) inst < 1.5 481  618.60 1 ( 0.34304 0.65696 ) *
     7) inst > 1.5 111  152.80 0 ( 0.54955 0.45045 )
      14) somadc < 199.5 88  115.40 0 ( 0.63636 0.36364 )
        28) Idade < 69.85 58   61.72 0 ( 0.77586 0.22414 ) *
        29) Idade > 69.85 30   39.43 1 ( 0.36667 0.63333 ) *
      15) somadc > 199.5 23   24.08 1 ( 0.21739 0.78261 )
        30) fa30 < 15.5 15    0.00 1 ( 0.00000 1.00000 ) *
        31) fa30 > 15.5 8   10.59 0 ( 0.62500 0.37500 ) *
```

```
> plot(arv.hipaf)
> text(arv.hipaf)
```

COMANDOS PARA A REGRESSÃO LOGÍSTICA

```
 ############################

 #Ajuste do Modelo VI
 pmod6 <-


dadff[-c(1,2,3,7,8,9,10,11,12,14,31,32,33,15,16,17,19,20,22,23,24,27,13,5,28,18,
 4)]
 mod6 <- glm(hipertensao ~
 .,family=binomial(link=~Tlogit~T),data=pmod6)

 #Sumário do ajuste do modelo, mostra as estimativas
 dos parâmetros e o erro padrão
 summary(mod6)

 #Análise de variância e teste de significância
 qui-quadrado para covariáveis
 anova(mod6, test=~SChisq~T)
```

```
 #Verificando o Modelo
 dadostest6<-


dadostest[-c(1,2,3,7,8,9,10,11,12,14,31,32,33,15,16,17,19,20,22,23,24,27,
+13,5,28,18,4)]
 preditosteste6<-predict(mod6,newdata=dadostest6)
 RL6Y <- dadostest6$hipertensao
 RL6A <- preditosteste6
 RL6A[RL6A>=0]<-1
 RL6A[RL6A<0]<-0

 class(RL6A)
 RL4A <- as.integer(RL6A)
 class(RL6A)

 MRL6 <- data.frame(RL6Y, RL6A)
 write.table(MRL6, " MRL6")
 table(RL6Y, RL6A)
```

COMANDOS R PARA REDES NEURAIS ARTIFICIAIS

```
> require(nnet)
> dadosrna1 <- dadostrein[c(13, 15, 16, 17, 18, 19, 20, 21, 22,
+     24, 25, 26, 28, 29)]
> padroniza = function(x) {
+     pad = (x - mean(x))/sd(x)
+ }
> dadosrna2 = as.matrix(dadosrna1)
> dadospad = apply(dadosrna2, 2, padroniza)
> colnames(dadospad) = names(dadosrna1)
> dadospad = data.frame(dadospad)
> dadospad$hipertensao = dadosrna1$hipertensao
> dadosrnat1 <- dadostest[c(13, 15, 16, 17, 18, 19, 20, 21, 22,
+     24, 25, 26, 28, 29)]
> dadosrnat2 = as.matrix(dadosrnat1)
> dadospadt = apply(dadosrnat2, 2, padroniza)
> colnames(dadospadt) = names(dadosrnat1)
> dadospadt = data.frame(dadospadt)
> dadospadt$hipertensao = dadosrnat1$hipertensao
```

```
> set.seed(45)
> rede = nnet(factor(hipertensao) ~ cabd + ccoxa + dcabd + dcsupra +
+     dctric + dccoxa + fa30 + lc30 + tc6x + Idade + imc + rcq +
+     exerc, data = dadospad, size = 20)


> preditos = predict(rede, dadospadt)
> preditos = round(preditos, 0)
> valid <- data.frame(dadostest$hipertensao, preditos)


> table(valid)
```