# Aplicação de técnicas de reconhecimento de padrões para a investigação de Síndrome de Down no primeiro trimestre de gravidez

# Application of pattern recognition techniques for the investigation of Down syndrome in pregnancy first trimester

**Leocir Bettiollo Junior**
Universidade Federal do Paraná – UFPR - PR
*bettiollo@gmail.com*

**Maria Teresinha Arns Steiner**
Universidade Federal do Paraná – UFPR - PR
*tere@ufpr.br*

**Resumo:** O presente trabalho tem por objetivo comparar o desempenho das técnicas de Redes Neurais Artificiais (RNAs) e de Regressão Logística (RL) na classificação de padrões apresentados pela Avaliação Bioquímica de Risco Fetal (ABRF). Os resultados para as RNAs, assim como para a RL, foram obtidos através do software STATGRAPHICS. Estas duas técnicas de Reconhecimento de Padrões foram bastante eficientes na tarefa de classificação dos padrões apresentados, sendo que as RNAs classificaram corretamente cerca de 93% dos padrões do conjunto de treinamento e cerca de 85% dos padrões do conjunto de teste. Estes percentuais para a técnica de RL foram de 93% e 86%, respectivamente.

**Palavras-chave:** redes neurais artificiais; regressão logística; síndrome de Down.

**Abstract:** The goal of this paper is to compare the performance of the Artificial Neural Networks (ANNs) and Logistic Regression (LR) techniques in the pattern classification presented by the Biochemical Evaluation Fetal Risk (BEFR). The results to the ANNs, as well as to the LR were obtained through STATGRAPHICS software. These two Pattern Recognition techniques were efficient in the pattern

classification task. The ANNs technique classified correctly about 93% of the patterns belonging to the training set and about 85% of the patterns belonging to the testing set. These values to the LR technique were 93% and 86%, respectively.

**Key words:** artificial neural networks; logistic regression; Down syndrome.

## 1 Introdução

Nos dias de hoje é muito comum mulheres terem filhos em idade mais avançada e este fato pode fazer com que ocorra um número muito superior de crianças com doenças genéticas cromossômicas.

Um dos tipos de doença genética cromossômica é a trissomia, que consiste na presença de três (e não dois, como seria normal) cromossomos de um tipo específico num organismo humano. Embora a trissomia possa ocorrer com qualquer cromossomo, os tipos mais comuns em humanos, são: Trissomia 21 (Síndrome de Down); Trissomia 18 (Síndrome de Edward); Trissomia 13 (Síndrome de Patau); Trissomia 8 (Síndrome de Warkany).

Muitas mulheres procuram especialistas para saber quais são as chances de ter uma gravidez com risco do feto ter (ou não) algum tipo de doença cromossômica e, para isso, podem ser realizados vários tipos de exames, sendo eles classificados como: invasivos ou não-invasivos.

Através de métodos invasivos podem-se gerar mapas cromossômicos (cariótipo fetal) que revelam, com certeza, a presença (ou não) de trissomias. A figura 1 apresenta uma ilustração de um cariótipo fetal com trissomia 21.

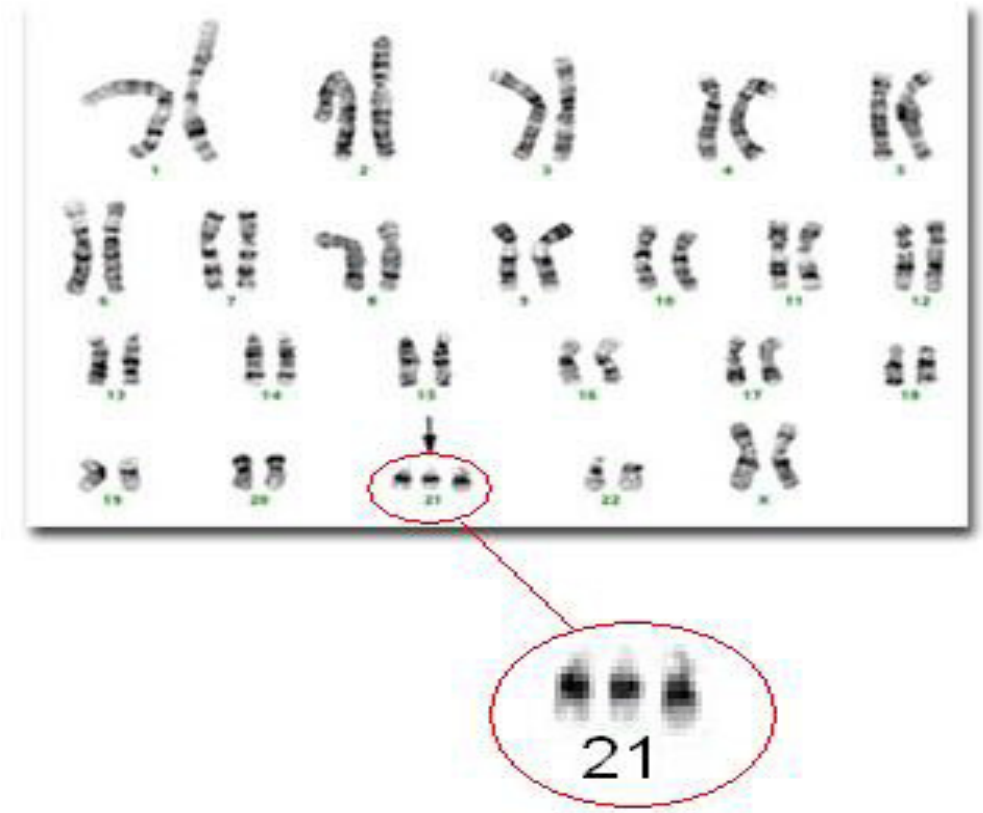


*Figura 1. Cariótipo fetal (trissomia 21)*

Na década de 70, a toda mulher gestante acima de 35 anos que procurasse o acompanhamento pré-natal eram ofertados exames invasivos como amniocentese (punção do líquido amniótico) ou biópsia de vilo (punção das vilosidades coriônicas)[1]. A cordocentese é outro método invasivo, onde é retirado, através de punção, sangue da veia do cordão umbilical. Estes métodos podem causar abortos espontâneos.

Existem métodos não invasivos capazes de fazer um rastreamento para auxiliar no diagnóstico da Síndrome de Down (SD), sem riscos de aborto. Uma combinação de exame de sangue com ultrassonografia pode diagnosticar a SD no feto durante o primeiro trimestre de gravidez[2]. Existe ainda um outro método não invasivo chamado Avaliação Bioquímica de Risco Fetal (ABRF) que consiste em analisar fetos com possíveis anormalidades cromossômicas. Essa análise identifica certos marcadores bioquímicos e os analisa de forma a obter um índice.

Este índice mostra se o risco do feto ter a SD é maior ou menor do que o risco de realizar alguns dos métodos invasivos citados anteriormente.

No primeiro trimestre de gravidez, o melhor marcador bioquímico é a dosagem conjunta da fração ß livre do hormônio gonadotrofina coriônica humana (HCG) e da Proteína Plasmática Associada à gravidez A (PAPP-A) [3]. Quando possível, este marcador é combinado com a medida da translucência nucal e a idade materna, atingindo uma elevada taxa de sensibilidade para detecção de fetos com SD (90%) [4]. Na proteína gonadotrofina coriônica humana (HCG), onde a mediana é 1,0 MoM (múltiplos da mediana), os níveis em casos de fetos com SD são altos com uma mediana de 1,63 MoM [5]. Valores acima de 1,63 são considerados preocupantes em relação à síndrome.

Estudos realizados nos anos 90 mostraram que a PAPP-A apresentou, em gestações de fetos anormais, dosagens significantemente baixas [6]. Ainda, de acordo com os relatos destes mesmos autores, o valor da mediana de PAPP-A é de 0,43 entre oito a quatorze semanas de gestação; valores abaixo de 0,43 são considerados muito significativos para prever se o feto terá a doença.

Em 1992, Nicolaides e Figueiredo cunharam o termo Transluscência Nucal (TN) [7] e avaliaram gestantes de dez a quatorze semanas gestacionais, considerando a TN anormal acima de três milímetros; o método demonstrou 67% para casos de SD. A idade materna (IM) é outro fator preocupante em relação à SD.

Para auxiliar no diagnóstico da possibilidade de uma criança nascer com SD, este presente trabalho busca aplicar os dados bioquímicos anteriormente citados, juntamente com a IM e a medida da TN, a duas técnicas de Reconhecimento de Padrões (RP) clássicas e já consagradas na literatura [8], utilizadas para a classificação de padrões: Redes Neurais Artificiais (RNAs) e Regressão Logística (RL). O objetivo é que tanto a técnica de RNAs quanto a de RL sejam capazes, após serem treinadas e testadas, identificar, com a máxima precisão, o nível do risco (alto ou baixo) do feto ser portador da doença, evitando-se a utilização dos métodos invasivos. Deve-se levar em consideração, ainda, que se faz necessária a utilização de métodos invasivos como a amniocentese (risco de 0,5% de aborto) e a biópsia de vilocorial (risco de 1% de aborto, podendo chegar a 2%) para chegar a uma análise conclusiva e satisfatória sobre o feto.

Este trabalho está estruturado da seguinte forma: na seção 2, são apresentados os dados coletados para o seu desenvolvimento, assim como a codificação dos dados. Na seção 3, são apresentados alguns trabalhos correlatos que fazem uso de técnicas de RP. Já, na seção 4, são apresentadas as duas técnicas de RP aqui utilizadas: RNAs e RL. A implementação de tais técnicas, assim como os resultados obtidos, são apresentados na seção 5, considerando-se os atributos originais e codificados. Finalmente, na seção 6 são apresentadas as conclusões e sugestões para trabalhos futuros.

## 2 Coleta e codificação de dados para o problema

Para o desenvolvimento do trabalho, foram coletados dados (padrões) de 450 mulheres que estiveram grávidas, das quais 120 (resposta "1") tiveram filhos com a SD e 330 tiveram filhos sem a SD (resposta "0"). Cada um destes padrões ficou constituído pelos quatro atributos (variáveis), já anteriormente descritos: HCG, PAPP-A, IM e TN. A saída (resposta) para cada um destes padrões pode assumir os valores "0", casos onde o risco do feto ter a SD é considerado inferior ao risco de aborto, caso seja realizado um exame invasivo e "1" para casos onde o risco é considerado superior.

As simulações realizadas através das duas técnicas de RP aqui utilizadas (RNAs e RL) fizeram uso dos dados "brutos", ou seja, dos dados originais e, também, dos dados na forma codificada, na tentativa de melhorar a eficiência das referidas técnicas. Tal codificação para cada um dos quatro atributos está explicitada nas tabelas de 1 a 4.

*Tabela 1. Codificação da variável Idade Materna*

| **Variável IM (Anos)** | **Codificação** |
|---|---|
| Até 20 | 1 |
| 20 a 30 (inclusive) | 2 |
| 30 a 40 (inclusive) | 3 |
| Acima de 40 | 4 |

*Tabela 2. Codificação da variável HCG*

| Variável HCG (MoM) | Codificação |
|---|---|
| Abaixo de 1,63 | 0 |
| 1,63 ou maior | 1 |

*Tabela 3. Codificação da variável TN*

| Variável TN (mm) | Codificação |
|---|---|
| Abaixo de 3 | 0 |
| 3 ou superior | 1 |

*Tabela 4. Codificação da variável PAPP-A*

| Variável PAPP-A(MoM) | Codificação |
|---|---|
| Acima de 0,43 | 0 |
| 0,43 ou inferior | 1 |

## 3 Trabalhos correlatos

São muitos os trabalhos na literatura que fazem uso de técnicas de RP aplicadas a problemas das mais diversas áreas, desde problemas médicos, de engenharia de avaliações, de indústrias, de crédito bancário, dentre outros. Dentre estes numerosos trabalhos, breves relatos de alguns deles são apresentados a seguir.

Em Baptistella et al. [9] é proposta a utilização das RNAs para a determinação dos valores venais de imóveis urbanos. Foram coletados 256 registros históricos (padrões) de imóveis urbanos da cidade de Guarapuava (PR). Cada um destes registros ficou composto por treze informações (atributos): bairro, setor, pavimentação, esgoto, iluminação pública, área do terreno, pedologia, topografia, situação, área edificada, tipo, estrutura e conservação. Várias simulações foram desenvolvidas, sendo que os piores resultados apresentaram acurácia de 78% e os melhores, de 95%.

Ainda, na avaliação imobiliária, tem-se ainda o que compara o desempenho das RNAs com a Análise de Regressão Múltipla para a venda de casas de

família[10]. Múltiplas comparações foram feitas entre os dois modelos nas quais foram variados: o tamanho da amostra de dados, a especificação funcional e a predição temporal. Em [11], os autores examinam o efeito que a vista de um lago (Lago Erie, E.U.A.) tem sobre o valor de uma casa. No estudo foram levados em consideração os preços baseados na transação das casas (preço de mercado). Os resultados indicam que, além da variável vista, que se apresenta significativamente mais importante do que as demais, também a área construída e o tamanho do lote são importantes.

Em [12], os autores utilizaram uma ferramenta baseada em Programação Genética para a classificação de proteínas, usando operadores especialmente projetados para o problema em pauta. O artigo apresenta um sistema para descobrir características que ocorrem frequentemente em proteínas de uma dada família, mas que raramente ocorrem em proteínas de outras famílias. Estas características poderão, então, ser usadas para classificar proteínas desconhecidas, predizendo suas funções. Experimentos foram desenvolvidos com um conjunto de enzimas extraído de um banco de dados de proteínas. Os resultados mostram que a técnica utilizada é muito eficiente em predizer características para a classificação de proteínas.

Já, em [13], os autores propõem a utilização de um modelo de Programação Linear (PL) para o RP de bobinas de papel de boa ou baixa qualidade. Foram coletados dados de 145 bobinas de papel (padrões), quarenta de boa qualidade e 105 de baixa qualidade. De cada bobina foram considerados dezoito atributos: testes de tração e rasgo da celulose, da pasta mecânica e da pasta termo-mecânica; quantidades destas três pastas; consistência e vazão da celulose e dados de sete rolos de prensagem da máquina de papel. A partir do modelo de PL, foi construído um segundo modelo matemático que faz uso do primeiro, de forma a garantir a obtenção de bobinas de boa qualidade a um mínimo custo.

A importância da análise exploratória dos dados preliminarmente à utilização das técnicas de RP é mostrada através de um problema médico [14]. As técnicas de RP utilizadas, comparativamente foram: Geração de uma Superfície que Minimiza Erros (PL); Função Discriminante Linear de Fisher; Modelo de RL; RNAs e Árvores de Decisão. O problema médico, de forma

simplificada, relacionava-se a discriminar pacientes com câncer ou cálculo no duto biliar. Foram considerados dados de 118 pacientes (padrões) do Hospital das Clínicas de Curitiba (PR), dos quais 35 possuíam câncer e 83, cálculo no duto biliar. De cada paciente foram considerados quatorze atributos: idade, sexo e doze resultados de exames laboratoriais. Os métodos de PL e de RNAs foram as técnicas que apresentaram maior acurácia dentre as técnicas abordadas, com cerca de 97% de acerto.

RNAs são utilizadas para o diagnóstico de cefaléia [15]. Para tanto, foram utilizados dados de 2.177 pacientes (padrões) e quatorze informações (atributos) de cada um deles: sexo, idade, além de dados sobre a dor: início, localização, intensidade, característica, surgimento, evolução, frequência, duração, fatores associados atenuantes e exacerbantes e uso de medicações. A acurácia das diversas simulações desenvolvidas variaram de 87% a 98%.

O algoritmo chamado Neurorule faz a extração de regras a partir de uma rede neural treinada, obtendo regras do tipo se-então [16, 17]. O desempenho desta abordagem é verificada, em ambos os artigos, em um problema de crédito bancário, sendo que a fim de facilitar a referida extração de regras, os valores dos atributos numéricos foram discretizados, dividindo-os em sub-intervalos. Após a discretização, o esquema de codificação termômetro foi empregado para obter representações binárias dos intervalos anteriormente definidos obtendo-se, assim, as entradas para a rede neural. Os resultados obtidos nos artigos indicam que, usando a abordagem proposta, regras de alta qualidade podem ser descobertas a partir de um conjunto de dados.

Fidelis et al. [18] apresentam um algoritmo de classificação baseado em Algoritmos Genéticos que descobre regras compreensíveis do tipo IF-THEN no contexto de Data Mining. O Algoritmo Genético proposto foi avaliado em duas bases de dados de domínio público, médicos de dermatologia e de câncer de mama, obtidos do UCI (*University of California at Irvi*ne) – Repositório de Aprendizado de Máquina (*Machine Learning Repository*).

Neurorule; Trepan e Nefclass são três métodos para a extração de regras de uma rede neural. Para comparar os desempenhos destes métodos foram utilizadas três bases de dados reais de crédito: German Credit (obtida do repositório UCI),

Bene 1 e Bene2 (obtidas das duas maiores instituições financeiras da Benelux)[19]. Os algoritmos mencionados são ainda comparados com os algoritmos C4.5-árvore, C4.5-regras e Regressão Logística Os autores ainda mostram como as regras extraídas podem ser visualizadas como uma tabela de decisão na forma de um gráfico compacto e intuitivo, permitindo uma melhor leitura e interpretação dos resultados ao gerente de crédito.

Na seção 4 é feita a descrição das técnicas de RP utilizadas neste trabalho para a predição da SD: RNAs e RL.

## 4 Técnicas utilizadas neste trabalho

Conforme já comentado, para o presente estudo foram abordadas duas técnicas de RP, RNAs e RL.

### 4.1 Redes neurais artificiais

Uma RNA é um processador paralelo e distribuído, constituído de unidades de processamento simples (neurônios), que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso[20].

Os neurônios que compõem uma RNA executam cálculos matemáticos que simulam o comportamento dos neurônios biológicos, fazendo com que essa rede possa aprender e se adaptar de acordo com a necessidade. A figura 2 apresenta o neurônio artificial, em que as entradas (ou atributos) $x_j$ (j = 1, ..., m), comparáveis aos dendritos dos neurônios biológicos, são as informações que serão passadas ao corpo da célula. Essas informações sofrerão alterações de acordo com os pesos $w_{kj}$ (k=1), já que o corpo da célula recebe o produto entre as entradas $x_j$ e os pesos $w_{kj}$.

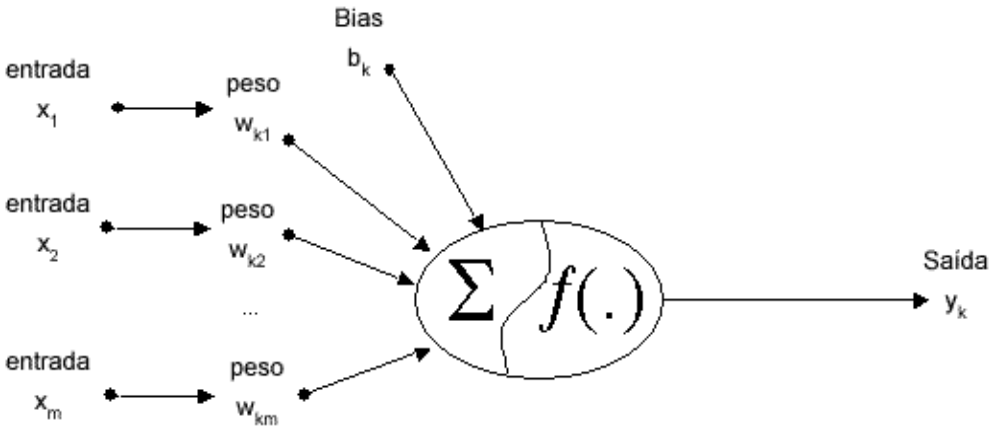


*Figura 2. Neurônio artificial*

O corpo da célula (soma) é comparável ao centro de processamento do neurônio artificial. Este processamento é formado por uma função soma, seguida por uma função de transferência. A função soma tem a função de somar os sinais de entrada, ponderados pelos respectivos pesos (sinapses); é representado pela equação (1). A função transferência tem a função de calcular a saída $y_k$ do neurônio, obtida de acordo com a função de transferência utilizada. Ainda, é inserido o bias $b_k$ que é uma entrada adicional ao neurônio artificial, sendo que seu valor de entrada é fixo em 1 e seu peso é ajustável como os demais pesos da rede.

$$V_k = \sum_{j=1}^{m} x_j w_{kj} \qquad (1)$$

A saída deste neurônio fará parte da resposta da rede, caso este neurônio se encontre na camada de saída ou será uma nova entrada para outro neurônio, caso este neurônio se encontre nas camadas intermediárias. Ao se combinar vários neurônios artificiais, tem-se formada uma rede de neurônios artificiais, ou seja, o objeto deste estudo, as RNAs. Existem inúmeras possibilidades de combinações de neurônios e, portanto, várias estruturas de redes neurais.

Em uma RNA do tipo *feed-forward*, em geral, trabalha-se com três camadas. A primeira camada é a camada de entrada, a qual recebe os atributos de cada um dos padrões que serão processados pelas camadas intermediárias, também chamadas de camadas ocultas. Nas camadas intermediárias é onde ocorre a maior

parte do processamento de uma RNA. Os dados são recebidos através das sinapses e, depois de processados, são enviados à camada de saída.

A camada de saída é responsável pela apresentação do resultado obtido pelo processamento da RNA. Os resultados obtidos pela rede são, então, comparados com os resultados desejados (0 ou 1, conforme definido anteriormente) e o erro é calculado. Com base no erro, os pesos entre os neurônios das três camadas são ajustados e o processo continua, até que uma regra de parada seja atendida. Para cada situação-problema, uma estrutura (topologia; arquitetura) de rede é modelada, dependendo da quantidade de neurônios na camada oculta, para que a rede tenha um aprendizado satisfatório. A quantidade de neurônios na camada oculta, em geral, é definida empiricamente, através de testes, que apontarão a melhor configuração. Para o problema em estudo têm-se quatro neurônios na camada de entrada (um para cada um dos quatro atributos), um número empírico de neurônio na camada oculta e um único neurônio na camada de saída.

Para o treinamento da rede *feed-forward* deste trabalho, foi utilizado o algoritmo *back-propagation* que consiste em duas fases: a primeira é chamada de propagação *forward* (para frente), a segunda denomina-se *backward* (para trás). Na propagação *forward*, como ilustrado na figura 3, um padrão é apresentado à camada inicial e flui em direção a última camada, passando pela camada intermediária. Ao chegar à última camada, é gerada uma resposta que é comparada com a saída desejada para este padrão e calcula-se o erro.

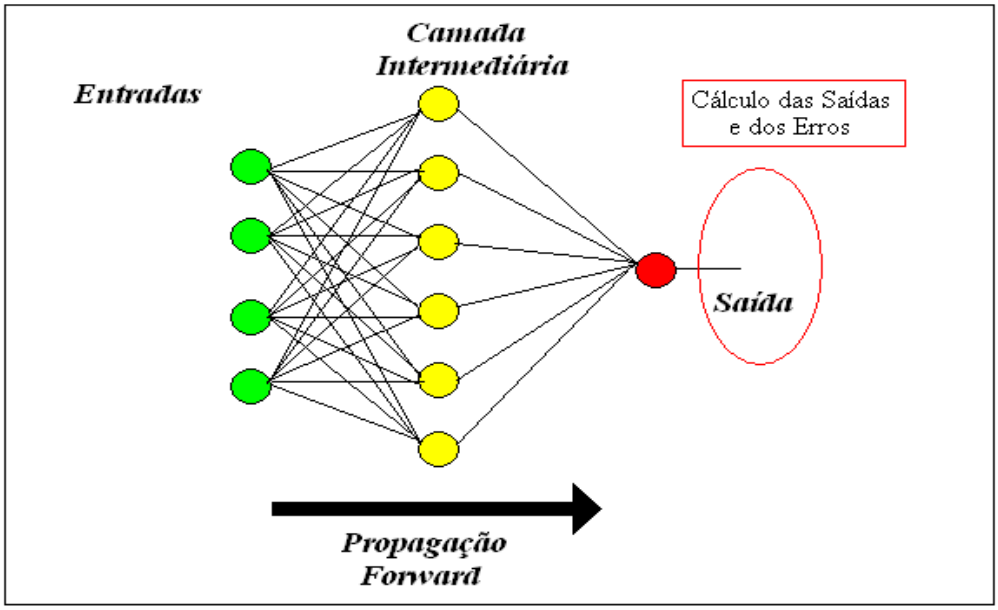


*Figura 3. Propagação forward*

Logo após o cálculo do erro, tem início a segunda etapa, a propagação backward, onde o erro é propagado a partir da camada de saída até a camada de entrada, de modo que os pesos das conexões vão sendo modificados, conforme o erro é retropropagado [21].

A figura 4 ilustra a propagação *backward*.

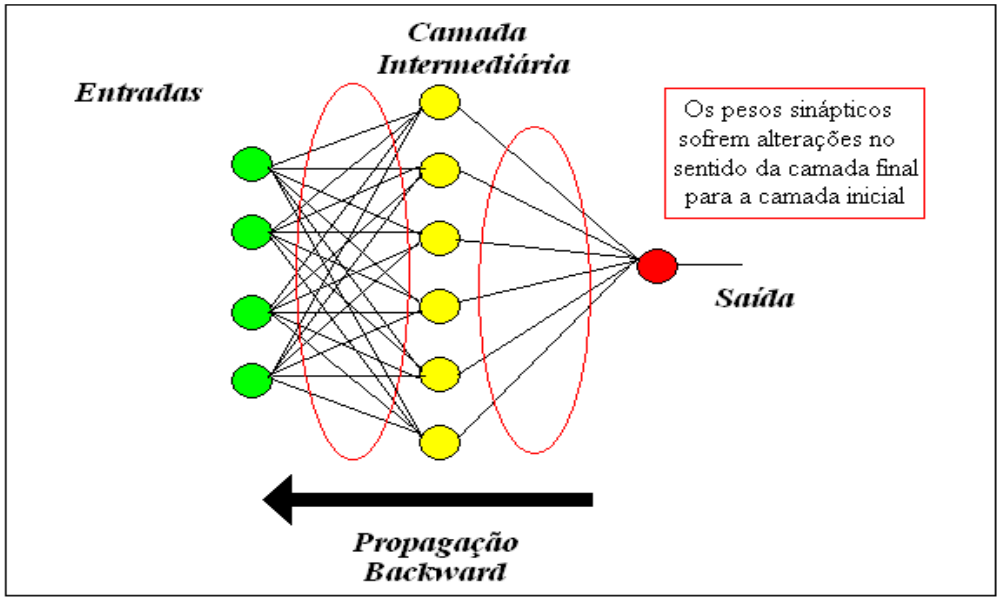


*Figura 4. Propagação backward*

## 4.2 Regressão logística

Muitas vezes, a posição dos pontos experimentais no diagrama de dispersão sugere a existência de uma relação funcional entre as duas variáveis [22]. Surge, então, o problema de determinar uma função que exprima esse relacionamento. Esse é o problema da regressão, conforme a denominação introduzida por Fisher, em 1936 e universalmente adotada [8].

Nas figuras 5 e 6 pode-se observar alguns pontos experimentais e, admitindo existir um relacionamento funcional entre os valores *x* e *y,* este relacionamento funcional recebe o nome de linha de regressão.

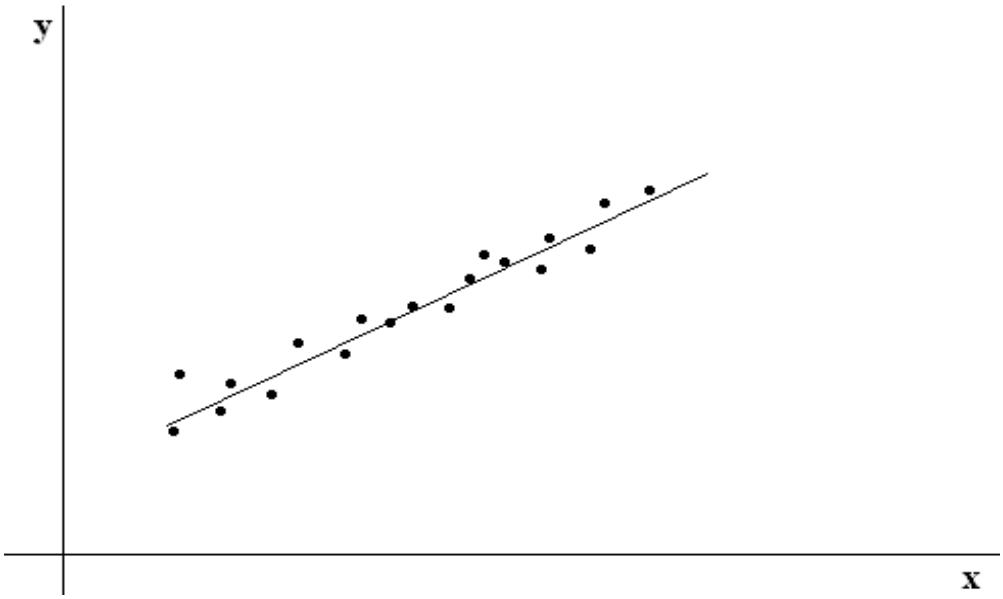


*Figura 5. Linha de regressão (linear simples)*

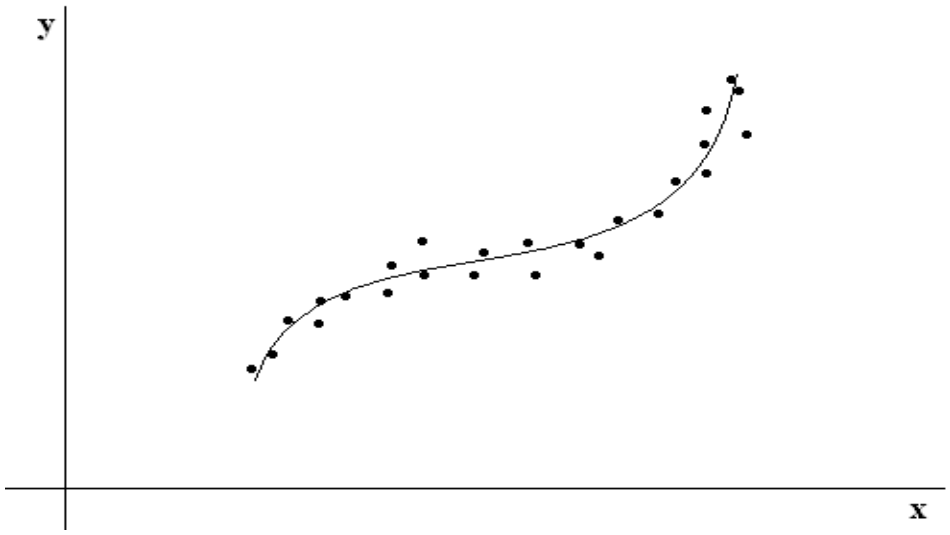


*Figura 6. Linha de regressão (polinomial)*

Estas linhas de regressão podem ser retas conhecidas por regressão linear simples ou linhas polinomiais, denotadas por regressão polinomial. Pode-se ter, ainda, a regressão linear múltipla, onde a variável dependente está relacionada com mais de uma variável. Existem, também, outros modelos de regressão. Para cada modelo de regressão, existe um função associada e o que se pretende encontrar são os coeficientes de cada função. Na tabela 5, são apresentados alguns modelos de regressão.

*Tabela 5. Alguns modelos de regressão*

| **Linear Simples** | **Polinomial (grau 2)** | **Linear Múltipla** |
|---|---|---|
| $y = \alpha + \beta x + \xi$ | $y = \alpha + \beta x + \gamma x^2 + \xi$ | $y = \alpha + \beta x_1 + \gamma x_2 + \xi$ |

No modelo de RL, usam-se os valores de uma série de variáveis independentes para predizer a ocorrência da variável dependente. Assim, todas as variáveis consideradas no modelo estão controladas entre si.

A RL, dentro da Análise Estatística Multivariada, consiste em relacionar, através de um modelo, uma variável resposta Y, dicotômica, com os fatores $(x_1, x_2, ..., x_{p-1})$ que influenciam as ocorrências de determinado evento [23].

### 4.2.1 Modelo logístico linear simples

Na RL estima-se diretamente a probabilidade de um evento ocorrer. Para um preditor (x), a probabilidade de um evento pode ser escrito como em (2).

$$\mathrm{Prob}(evento) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad (2)$$

Onde $\beta_0$ e $\beta_1$ são os coeficientes de regressão estimados e *e* é a base dos logaritmos naturais. O modelo logístico linear simples é derivado da função matemática em (3).

$$f(y) = \frac{1}{1+e^{-y}}, \quad y \in R \qquad (3)$$

multiplicando a parte superior e a inferior de f(y) por $e^y$, tem-se

$$f(y) = \frac{e^y}{e^y + 1} \qquad (4)$$

que é o Modelo Logístico Linear Simples.

### 4.2.2 Modelo logístico linear múltiplo

O modelo logístico linear múltiplo decorre da existência de vários preditores $(x_1, x_2, ..., x_k)$, ou seja, a variável dependente estará associada a várias variáveis independentes. A probabilidade de um evento pode ser escrito como em (5).

$$\mathrm{Prob}(evento) = \frac{1}{1+e^{-z}} \qquad (5)$$

onde z é uma combinação linear: $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ e os $\beta_n$, n = 1,2,...,k são os coeficientes de regressão estimados e *e* é a base dos logaritmos naturais.

## 5 Aplicação das técnicas e obtenção dos resultados

Para a aplicação das duas técnicas utilizadas neste trabalho (RNAs e RL), os 450 padrões, dos quais 120 relacionados à síndrome (resposta 1) e 330 não

relacionados à síndrome (resposta 0), conforme já descrito na seção 2, foram divididos em dois conjuntos: um conjunto para o treinamento destas técnicas com 300 padrões (73 com resposta 1 e 227 com resposta 0) e o outro para o teste com tais técnicas, com 150 padrões (47 com resposta 1 e 103 com resposta 0).

Algumas simulações foram realizadas com os atributos originais e outras com os atributos codificados, conforme tabelas 1 a 4, apresentadas na seção 2, fazendo-se uso do software estatístico STATGRAPHICS.

A apresentação dos resultados está de acordo com os dados utilizados: primeiramente, são reportados os resultados obtidos a partir dos dados originais e, em seguida, são apresentados os resultados para os dados codificados, tanto para as RNAs quanto para a RL.

## 5.1 Utilizando os dados originais

### 5.1.1 Redes neurais artificiais

O software STATGRAPHICS gerou, automaticamente, uma RNA com 10 neurônios na camada intermediária, como mostra a figura 7. Tal RNA apresenta 93% de acertos para o conjunto de treinamento, como mostra a tabela 6.

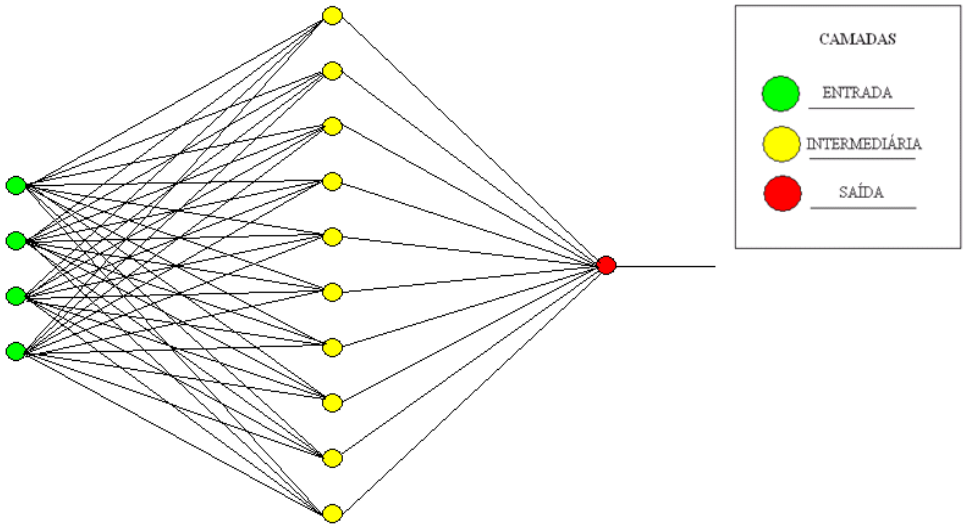


*Figura 7. Rede gerada pelo software STATGRAPHICS*

*Tabela 6. Desempenho da RNA para o conjunto de treinamento (dados originais)*

| Saída | Caso | Acertos | % de acertos |
|---|---|---|---|
| 0 | 227 | 222 | 97,7974 |
| 1 | 73 | 57 | 78,0822 |
| Total | 300 | 279 | 93,0000 |

Em relação ao conjunto de testes, a RNA, com os pesos já definidos no decorrer do treinamento, classificou corretamente 84% dos padrões, como mostra a tabela 7.

*Tabela 7. Desempenho da RNA para o conjunto de teste (dados originais)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 103 | 87 | 84,4660 |
| 1 | 47 | 39 | 82,9787 |
| Total | 150 | 126 | 84,0000 |

### 5.1.2 Regressão logística

A inserção dos dados e uso do software STATGRAPHICS para a RL resultaram em um modelo com a forma apresentada em (6).

$$f(y) = \frac{e^{y}}{e^{y} + 1} \tag{6}$$

onde:

y = -18,5035 + 1,17666*HCG + 0,309142*IM - 0,739673*PAPP-A + 2,83482*TN.

As tabelas 8 e 9 apresentam o desempenho deste modelo para os conjuntos de treinamento e testes, respectivamente.

*Tabela 8. Desempenho da RL para o conjunto de treinamento (dados originais)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 227 | 215 | 94,7136 |
| 1 | 73 | 64 | 87,6712 |
| Total | 300 | 279 | 93,0000 |

*Tabela 9. Desempenho da RL para o conjunto de teste (dados originais)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 103 | 74 | 71,8446 |
| 1 | 47 | 44 | 93,6170 |
| Total | 150 | 118 | 78,6667 |

## 5.2 Utilizando os dados codificados

### 5.2.1 Redes neurais artificiais:

Com os dados codificados, os resultados obtidos foram similares aos obtidos com os dados originais, dez neurônios na camada intermediária e apresentou uma acurácia de 93% para os padrões contidos no conjunto de treinamento, conforme apresentado na tabela 10.

*Tabela 10. Desempenho da RNA para o conjunto de treinamento (dados codificados)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 227 | 222 | 97,7974 |
| 1 | 73 | 57 | 78,0822 |
| Total | 300 | 279 | 93,0000 |

Para o conjunto de teste, a acurácia foi de 79,3333% dentre os 150 padrões aplicados a RNA gerada pelo software STATGRAPHICS. A tabela 11 mostra os acertos para cada tipo de saída e também no total.

*Tabela 11. Desempenho da RL para o conjunto de teste (dados codificados)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 103 | 76 | 73,7864 |
| 1 | 47 | 43 | 91,4893 |
| Total | 150 | 119 | 79,3333 |

### 5.2.2 Regressão logística

Já o modelo logístico (7), com os dados codificados, classificou corretamente 91,6667% dos padrões do conjunto de treinamento, conforme mostra a tabela 12.

$$f(y) = \frac{e^y}{e^y + 1} \quad (7)$$

onde:

y =-8,27157 + 5,17552*HCG + 1,82787*IM + 1,35433*PAPP-A + 1,48458*TN.

*Tabela 12. Desempenho da RL para o conjunto de treinamento (dados codificados)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 227 | 218 | 96,0352 |
| 1 | 73 | 57 | 78,0821 |
| Total | 300 | 275 | 91,6667 |

Para o conjunto de teste o resultado obtido foi de 86% de acertos, conforme apresentado na tabela 13.

*Tabela 13. Desempenho da RL para o conjunto de teste (dados codificados)*

| Saída | Casos | Acertos | % de acertos |
|---|---|---|---|
| 0 | 103 | 87 | 84,4660 |
| 1 | 47 | 42 | 89,3617 |
| Total | 150 | 129 | 86,0000 |

## 6 Conclusões e sugestões para trabalhos futuros

O presente trabalho fornece uma forma alternativa para auxiliar o diagnóstico de um feto ter a SD (ou não), através do uso de técnicas de RP como, por exemplo, as RNAs e a RL, sem a necessidade de fazer uso, portanto, de métodos invasivos, altamente prejudiciáveis à saúde da gestante. Estas técnicas conseguem aprender através de dados históricos de gestantes com filhos com SD (ou não), investigando quatro variáveis destas mulheres quando gestantes: HCG (dosagem conjunta da fração ß livre do hormônio gonadotrofina coriônica humana), PAPP-A (Proteína Plasmática Associada à gravidez A), IM (Idade Materna) e TN (Transluscência Nucal). Para o desenvolvimento deste trabalho, foram coletados um total de 450

dados (padrões, gestantes), dos quais 120 referem-se a crianças com SD e 330 sem SD. Os percentuais de acertos (acurácia) para as diversas simulações desenvolvidas, para ambas as técnicas, encontram-se resumidos na tabela 14.

*Tabela 14. Resumo dos procedimentos executados*

| | | | | **Redes Neurais (STATGRAPHICS)** | | |
|---|---|---|---|---|---|---|
| | **Quantidade** | **Saída 0** | **Saída 1** | **Acertos Saída 0** | **Acertos Saída 1** | **% de acertos** |
| Dados Originais | | | | | | |
| Treinamento | 300 | 227 | 73 | 222 | 57 | 93,0000% |
| Teste | 150 | 103 | 47 | 87 | 39 | 84,0000% |
| Dados Codificados | | | | | | |
| Treinamento | 300 | 227 | 73 | 222 | 57 | 93,0000% |
| Teste | 150 | 103 | 47 | 76 | 43 | 79,3333% |
| | **Quantidade** | **Saída 0** | **Saída 1** | **Regressão Logística (STATGRAPHICS) Acertos Saída 0** | **Acertos Saída 1** | **% de acertos** |
| Dados Originais | | | | | | |
| Treinamento | 300 | 227 | 73 | 215 | 64 | 93,0000% |
| Teste | 150 | 103 | 47 | 74 | 44 | 78,6667% |
| Dados Codificados | | | | | | |
| Treinamento | 300 | 227 | 73 | 218 | 57 | 91,6667% |
| Teste | 150 | 103 | 47 | 87 | 42 | 86,0000% |

Tanto a técnica de RNAs quanto a de RL foram capazes de, após serem treinadas e testadas, identificar o nível do risco (alto ou baixo) do feto ser portador da doença. Os desempenhos das duas técnicas foram análogos, sendo que a melhor acurácia para as RNAs foi obtida ao se trabalhar com os dados originais (84%) e para a RL, ao se trabalhar com os dados codificados (86%), considerando-se os resultados para o conjunto de testes que aponta a capacidade de generalização da técnica para ambos os casos. Assim, dado um novo caso (gestante) para o qual se deseja conhecer tal risco, pode-se aplicar a RNA com os dados originais da gestante à RNA com 10 neurônios na camada oculta e pesos já definidos no treinamento ou, então, a RL com os dados codificados da gestante, levando-se em conta a equação particularizada para (7).

Deve-se levar em consideração, ainda, que se faz necessária a utilização de métodos invasivos como a amniocentese (risco de 0,5% de aborto) e a biópsia de vilocorial (risco de 1% de aborto, podendo chegar a 2%) para se chegar a uma análise conclusiva e satisfatória sobre o feto com 100% de acerto.

Sugere-se para trabalhos futuros, a aplicação de outras técnicas de RP como, por exemplo, outras técnicas clássicas e/ou metaheurísticas, além de formas alternativas de análise exploratória dos atributos, preliminarmente a aplicação das técnicas, sempre tendo em vista a maximização da acurácia.

## 7 Referências

[1] SIMPSON, J. L. Choosing the best prenatal screening protocol. *New Engl J Med* v. 353, n. 19, p. 2068-2070, 2005.

[2] PUPO FILHO, R. A. *Síndrome de Down* - Causada por uma Alteração Cromossômica, é o Tipo Mais Comum de Retardo Mental. São Paulo, 2000. Disponível em: <http://boasaude.uol.com.br/lib/ShowDoc.cfm?LibDocID=3789& ReturnCatID=1798>. Acesso: 15 jan. 2008.

[3] HADDOW, J. E. et al. Screening of maternal serum for fetal Down´s syndrome in the first trimester. *New Engl J Med* v.338, p. 955-961, 1998.

[4] SOERGEL P. et al. Screening for trisomy 21 with maternal age, fetal nuchal translucency and maternal serum biochemistry at 11-14 weeks: a regional experience from Germany. *Fetal Diagn Ther* v. 21, n. 3, p. 264-268, 2006.

[5] CASALS, E. et al. First-trimester biochemical markers for down syndrome. *Prenatal Diag* v. 19, p. 08-11, 1999.

[6] BERSINGER, N. A. et al. Production and characterization of monoclonal antibodies against pregnancy – associated plasma protein A. *Mol Hum Reprod* v. 5, n. 7, p. 675-681, 1999.

[7] NICOLAIDES, K. H.; FIGUEIREDO, D. B. *O exame ultra-sonográfico entre 11-13+6 semanas*. Fetal Medicine Foundation, London, 2004.

[8] JOHNSON, R. A.; WICHERN, D. W. Applied Multivariate Statistical Analysis. *Prentice Hall*, inc., 4. ed., Nova Jersey, 1998.

[9] BAPTISTELLA, M.; CUNICO, L. H. B.; STEINER, M. T. A. O Uso de Redes Neurais na Engenharia de Avaliações: Determinação dos Valores Venais de Imóveis Urbanos, *Rev Cienc Exatas Nat* v. 9, n. 2, p. 215-229, 2009.

[10] NGUYEN, N.; CRIPPS, A. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *J Real Estate Res* v. 22, n. 3, p. 313-336, 2001.

[11] BOND, M.T.; SEILER, V. L.; SEILER, M. J. Residencial Real Estate Prices: a Room with a View. *J Real Estate Res* v. 23, n. 1, p. 129-137, 2002.

[12] TSUNODA, D. F.; FREITAS, A. A.; LOPES, H. L. MAHATMA: A Genetic Programming-Base Tool for Protein Classification. International Conference on Intelligent Systems Design and Applications Conference (ISDA), Pisa, Italy, 30/nov-02/dez, 2009.

[13] STEINER, M. T. A.; CARNIERI, C.; STANGE, P. Construção de um Modelo Matemático para o Controle do Processo de Produção do Papel Industrial. *Pesq Oper Desenvolv* v. 1, n. 1, p. 33-49, 2009.

[14] STEINER, M. T. A. et al. Abordagem de um Problema Médico por meio do Processo de KDD com ênfase à Análise Exploratória dos Dados. *Rev Gest Prod* v. 13, n. 2, p. 325-337, 2006.

[15] MENDES, K. B.; STEINER, M. T. A. Diagnóstico de Dor de Cabeça usando Redes Neurais Artificiais. *Rev IST* v. 8, p. 41-47, 2008.

[16] LU, H.; SETIONO, R.; LIU, H. NeuroRule: A Connectionist Approach to Data Mining. *Proceedings of the 21st. VLDB Conference*, Switzerland, p. 478-489, 1995.

[17] LU, H.; SETIONO, R.; LIU, H. Effective Data Mining using Neural Networks. *IEEE T Knowl Data En* v. 8, n. 6, p. 957-961, 1996.

[18] FIDELIS, M.V.; LOPES, H.S.; FREITAS, A.A. Um Algoritmo Genético para Descobrir Regras de Classificação em Data Mining. *Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação*, v. IV, p. 17-29, 2000.

[19] BAESENS, B.; SETIONO, R.; MUES, C.; VANTHIENEN, J. Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Manage Sci* v. 49, n. 3, p. 312-329, 2003.

[20] HAYKIN, S. *Redes Neurais* - Princípios e Prática. Bookman, 2. ed., Porto Alegre, RS, 2001.

[21]CARVALHO, A. P. L. F. *Redes Neurais Artificiais*, 2000. Disponível em:< http://www.icmc.sc.usp.br/~andre/>. Acesso: 25 jul. 2009.

[22] COSTA NETO, P. L. O. *Estatística*. 2. ed. São Paulo: Edgard Blücher 2002.

[23] CHAVES NETO, A. *Notas de Aulas Apresentadas na disciplina de Análise Multivariada Aplicada à Pesquisa*, PPGMNE, UFPR, Curitiba, PR, 2008.