REBECA NONATO SILVA

IDENTIFICAÇÃO DE PACIENTES COM POTENCIAL PARA DESENVOLVER O PÉ DIABÉTICO BASEADA EM TÉCNICAS DE RECONHECIMENTO DE PADRÕES E AÇÕES DE AUTOCUIDADO

LAVRAS - MG

2014

REBECA NONATO SILVA

IDENTIFICAÇÃO DE PACIENTES COM POTENCIAL PARA DESENVOLVER O PÉ DIABÉTICO BASEADA EM TÉCNICAS DE RECONHECIMENTO DE PADRÕES E AÇÕES DE AUTOCUIDADO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Modelagem de Sistemas Biológicos, para a obtenção do título de Mestre.

Orientador

Dr. Danton Diego Ferreira

Coorientador

Dr. Bruno Henrique Groenner Barbosa

LAVRAS - MG

2014

Ficha Catalográfica Elaborada pela Coordenadoria de Produtos e Serviços da Biblioteca Universitária da UFLA

Silva, Rebeca Nonato.
Identificação de pacientes com potencial para desenvolver o pé diabético baseada em técnicas de reconhecimento de padrões e ações de auto cuidado / Rebeca Nonato Silva. – Lavras : UFLA, 2014.
84 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2014.
Orientador: Danton Diego Ferreira.
Bibliografia.

1. Pé diabético. 2. K-means. 3. Autocuidado. 4. Sistemas Biológicos – simulação. I. Universidade Federal de Lavras. II. Título.

CDD – 570.11

REBECA NONATO SILVA

IDENTIFICAÇÃO DE PACIENTES COM POTENCIAL PARA DESENVOLVER O PÉ DIABÉTICO BASEADA EM TÉCNICAS DE RECONHECIMENTO DE PADRÕES E AÇÕES DE AUTOCUIDADO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, área de concentração em Modelagem de Sistemas Biológicos, para a obtenção do título de Mestre.

APROVADA em 02 de julho de 2014.

Dr. Bruno Henrique Groenner Barbosa	UFLA
Dr. Vilma Elenice Contatto Rossi	FESPMG

Dr. Danton Diego Ferreira
Orientador

LAVRAS - MG

2014

AGRADECIMENTOS

Durante o tempo de realização deste Mestrado, muitas pessoas contribuíram em meus trabalhos das mais diferentes formas. A todos, os meus sinceros agradecimentos.

Agradeço à Universidade Federal de Lavras – UFLA, ao Departamento de Engenharia - DEG e ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação - PPGESISA, pela estrutura oferecida e pela oportunidade de realização do Mestrado.

Agradeço a Capes pela concessão da bolsa de estudos, que tornou possível a realização do Mestrado.

Em especial ao meu Orientador e Co-orientador por todo auxílio durante a jornada.

RESUMO

O pé diabético é uma das complicações mais graves do *diabetes mellitus*. Geralmente, a partir de uma úlcera no pé, pode ocorrer amputação de uma ou toda extremidade inferior se a úlcera não for adequadamente tratada. Com o propósito de evitar o pé diabético, este trabalho propõe um sistema automático não invasivo com base em técnicas de inteligência computacional e reconhecimento de padrões para identificar pacientes com diabetes que apresentam alto risco de desenvolver o pé diabético. Para projetar o método, foram recolhidas informações relativas ao âmbito social, hábitos e autocuidados dos pacientes diabéticos. Os dados foram agrupados em dois e três grupos utilizando-se o algoritmo *K-means*. Em seguida, os classificadores baseados nos centroides dos grupos obtidos, em redes neurais, árvores de decisão e no K vizinhos mais próximos, foram utilizados, comparativamente para classificar os pacientes diabéticos em alto ou baixo risco de desenvolver o pé diabético. Ambos os dados reais e simulados foram usados para avaliação do método. Desempenhos de 100% para dados simulados e 68%, considerando-se a classificação dos especialistas como o padrão-ouro para dados reais, foram obtidos. O método requer um processamento computacional simples e pode ser útil para Unidades Básicas de Saúde para triagem de pacientes diabéticos ajudando a equipe de saúde a reduzir o número de casos de pé diabético.

Palavras-chave: Pé diabético. *K-means.* Autocuidado.

ABSTRACT

The diabetic foot is one of the most serious complications of diabetes mellitus. Generally, from a foot ulcer, amputation of a lower extremity or all may occur. In order to avoid diabetic foot, this work proposes an automatic non invasive system based on computational intelligence algorithms to identify patients with diabetes who have a high risk of developing diabetic foot. To design the method, information was collected regarding the social, habits and self-care of diabetic patients. We used the K-means algorithm to divide the data into two and three groups. After this, classifiers based on centroids of the groups, neural networks, decision tree and k-neighbor nearest were applied to classify the diabetic patient as being of high or low risk of developing diabetic foot. Both real and simulated data were used to evaluate the method. Accuracy of 100% for simulated data and 68% for real data, considering the classification of experts as the gold standard were achieved. The method requires a simple computational processing and can be useful for Basic Health Screening for diabetic patients to help the healthcare team to reduce the number of cases of diabetic foot.

Keywords: Diabetic foot. *K-means*. Self-Care.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A área da saúde vem se destacando na busca de alternativas para o auxílio ao diagnóstico médico. Para suprir essas alternativas, vê-se necessário o desenvolvimento de sistemas rápidos e precisos no processamento de informações, ou seja, na Mineração de Dados. A Mineração de Dados vem conquistando seu espaço nos diagnósticos médicos, auxiliando na tomada de decisões dos profissionais de saúde. É desejável que os sistemas de apoio à triagem e diagnóstico médico sejam eficientes na detecção ou na identificação do grau de risco de se desenvolver uma dada doença, além de ser aliada a uma baixa incidência de falsos alarmes.

Os métodos de identificação de estados patológicos em pacientes são, geralmente, invasivos, dependentes de exames clínicos e levam algum tempo para que respostas definitivas sejam obtidas, o que é crítico no caso de doenças contagiosas. Com isso, há uma tendência em se utilizar métodos alternativos não invasivos para auxiliar o diagnóstico e fazer a triagem dos pacientes. Geralmente, estes métodos alternativos são baseados em inteligência computacional, reconhecimento de padrões e processamento de sinais, que a partir de um conjunto de informações acerca do paciente, oferecem uma resposta objetiva sobre seu estado patológico acompanhada de certo grau de certeza.

Diabetes mellitus é considerado um problema sério de saúde, sendo a quarta causa de morte no Brasil. A falta de insulina ou a incapacidade desta de exercer com suas funções provocam a síndrome de etiologia múltipla que é o *diabetes mellitus*. Em longo prazo, pode gerar falência múltipla nos órgãos chegando à morte. As lesões nas extremidades inferiores são frequentes na população diabética. Ao procurar um atendimento médico, se as lesões estão em estado avançado acabam requerendo uma cirurgia. Infecções acometendo o pé

são as principais infecções de partes moles em diabéticos. Tais infecções também são a principal causa de amputação.

Especialistas ressaltam a atenção especial com os pacientes diabéticos, uma vez que nem sempre a neuropatia diabética manifesta sintomas e um diagnóstico precoce pode ser a chave para evitar danos aos pés. Ainda é enfatizado que o quadro clínico do paciente vai piorando com o tempo: quando há um diagnóstico precoce, o controle glicêmico pode reverter a situação.

1.1 Justificativa e relevância do estudo

O manejo do diabetes não constitui uma tarefa fácil e deve ser feito dentro de um sistema hierarquizado de saúde, sendo sua base o nível primário.

Considerando-se a especificidade do diabetes como doença crônica e o controle glicêmico como fundamental na prevenção de complicações e sequelas, o conhecimento da doença por meio de informações e educação constitui aspecto extremamente relevante no tratamento (FERRAZ et al., 2000).

De acordo com Torres et al. (2009), o aumento da prevalência do diabetes aliado à complexidade de seu tratamento, tais como: restrições dietéticas, uso de medicamentos e complicações crônicas associadas (retinopatia, nefropatia, neuropatia, cardiopatia, pé neuropático, entre outras) reforçam a necessidade de programas educativos eficazes e viáveis aos serviços públicos de saúde.

As recomendações para o controle domiciliar do diabético incluem automonitorização: da glicemia capilar, de múltiplas doses de insulina, das alterações nos padrões dietéticos a partir de reeducação alimentar e da realização de atividades físicas programadas, a fim de manter os níveis glicêmicos (SANTOS; ROSSI; OLIVEIRA, 2011). Estas recomendações implicarão em mudanças de comportamento dos pacientes, profissionais de saúde e familiares.

De fato, quando tais recomendações são seguidas a risca, o controle glicêmico é possível evitando assim complicações.

No entanto, o sistema atual de saúde pública mostra-se ultrapassado e ineficaz ao tentar transmitir aos usuários as recomendações sobre o autocuidado, principalmente, quando os usuários são de baixo nível socioeconômico, devido a dificuldades em absorver as informações transmitidas. Tais usuários requerem atenção especial durante o acolhimento à unidade de saúde e, portanto, demandam mais esforços por parte do sistema público de saúde.

Santos, Rossi e Nascimento (2010) e Ferreira, et al. (2011) levantaram informações importantes sobre o grau de conhecimento e medidas de autocuidado adotadas por usuários de uma unidade de saúde da família em relação ao pé diabético. O estudo mostrou que os pacientes diabéticos, em geral, possuem dúvidas e dificuldades para entender todo o contexto de sua doença e suas complicações. Um dado mais alarmante foi o fato de que 57,69% dos pacientes portadores de diabetes entrevistados informaram não saber como prevenir o pé diabético. Este trabalho mostra a ineficiência do sistema público de saúde em desenvolver a educação continuada com relação ao diabetes aos pacientes.

No Brasil, apesar das várias experiências municipais bem sucedidas de acompanhamento dos casos de *diabetes mellitus*, observa-se, em grande parte do país, a falta de vínculos entre os pacientes e as Unidades de Saúde. Em geral, o atendimento aos pacientes ocorre de modo assistemático, nos serviços de emergência, espaço esse que não propicia nem a identificação de lesões em órgãos-alvo do diabetes, nem o estabelecimento de um programa de tratamento e acompanhamento adequado a cada caso. No que se refere à promoção da saúde e à prevenção dos fatores de risco, a situação é ainda mais crítica, tendo em conta a falta de preparação e de tradição dos serviços de saúde para a realização sistemática de tais atividades.

Além disso, a identificação oportuna de casos favorece o estabelecimento de vínculos entre os pacientes e as Unidades Básicas de Saúde, o que é imprescindível para o sucesso do controle da doença. Entre as vantagens do acompanhamento e controle do diabetes no âmbito da atenção básica está a possibilidade de evitar o surgimento de complicações, assim como de evitar o agravamento destas ocorrências, reduzindo tanto o número de internações hospitalares quanto a mortalidade relacionada às doenças cardiovasculares, entre outras complicações.

É consenso geral entre endocrinologistas e infectologistas que a principal medida no tratamento do pé diabético é a detecção precoce, alcançando-se mais de 90% de sucesso para as úlceras que recebem manejo adequado, incluindo alívio da pressão local, tratamento das infecções e revascularização quando indicada (TORRES et al., 2009).

Esta situação demonstra a necessidade de os serviços de saúde pública reverem suas práticas, com a implantação de ações para estabelecer medidas de prevenção e controle desta doença, com o objetivo de reduzir os índices de morbimortalidade (TORRES et al., 2009). Neste mesmo sentido, as orientações do Ministério da Saúde são para se efetuar intervenções em saúde antes da manifestação dos fenômenos patológicos. Assim, são recomendadas medidas preventivas, extrapolando-se as ações assistenciais e demandando-se práticas de saúde mais abrangentes para a população, a fim de minimizar o aparecimento dos fatores de risco ou reduzir a oportunidade de exposição das pessoas a eles.

Dentro deste contexto, o desenvolvimento de um sistema automático não invasivo de identificação de pacientes diabéticos com potencial para desenvolver o pé diabético, de forma a apoiar a equipe profissional na tarefa de detectar futuros pés diabéticos e, assim, efetuar um controle mais dinâmico do problema, fazendo-se a triagem dos pacientes cadastrados na Unidade Básica de Saúde é de extrema importância e constitui-se no objetivo deste trabalho.

Adicionalmente, o desenvolvimento deste trabalho contribuirá para a implantação de um plano de cuidado específico para cada tipo de paciente, baseado no grupo de risco em que cada um se encaixa, levando-se em consideração sua situação socioeconômica, hábitos e cuidados com os pés.

1.2 Objetivo

A presente dissertação teve como objetivo desenvolver um sistema automático não invasivo utilizando Inteligência Computacional para a identificação precoce de pacientes diabéticos com potencial para desenvolver o pé diabético.

Os objetivos específicos foram:

a) Construir um banco de dados, a partir da coleta de informações junto à comunidade de pacientes portadores do diabetes cadastrados nas Unidades Básicas de Saúde em uma cidade no interior de Minas Gerais utilizando o questionário específico mostrado no Anexo A; Foram levadas em consideração apenas as informações acerca do âmbito social, hábitos e do autocuidado com os pés dos pacientes. A ideia é verificar se tais informações podem modelar um sistema de identificação de pacientes em risco.
b) Implementar e comparar diferentes algoritmos de agrupamento de dados para definir grupos de risco presentes;
c) Desenvolver um modelo automático, a partir dos agrupamentos obtidos, para classificar os pacientes em graus de risco de desenvolver o pé diabético: i) alto risco e ii) baixo risco;
d) Desenvolver técnicas para seleção das variáveis mais discriminantes a cerca do desenvolvimento do pé diabético.

1.3 Organização do trabalho

A dissertação está organizada da seguinte forma. Na primeira, seção é apresentada a introdução, a relevância do estudo proposto e seus objetivos gerais e específicos. A próxima seção apresenta o Referencial Teórico, no qual serão introduzidas as ferramentas utilizadas ao longo do trabalho e o estado da arte da complicação pé diabético.

A terceira seção irá descrever os passos do trabalho, os procedimentos metodológicos com relação à pesquisa e os sujeitos envolvidos. É apresentado a base de dados e o projeto do método, além das fases operacionais. Na quarta seção, os resultados são apresentados, uma análise comparativa entre os principais resultados e os métodos utilizados, também é apresentada a classificação da especialista. E por fim, as conclusões obtidas com o trabalho proposto e os trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 *Diabetes mellitus*

O *diabetes mellitus* constitui, atualmente, um dos principais problemas de saúde, no que se refere tanto ao número de pessoas afetadas, gerando incapacidade e mortalidade, quanto ao elevado investimento do governo para o controle e tratamento de suas complicações (VIGO; NUNES; PACE, 2003). Ele já é a quarta causa de morte no Brasil.

O *diabetes mellitus* é uma síndrome de etiologia múltipla, decorrente da falta de insulina e/ou da incapacidade de a insulina exercer adequadamente seus efeitos. Caracteriza-se por hiperglicemia crônica com distúrbios do metabolismo dos carboidratos, lipídios e proteínas. As consequências do *diabetes mellitus*, em longo prazo, incluem disfunção e falência de vários órgãos, especialmente os rins, olhos, nervos, coração e vasos sanguíneos (BRASIL, 2002).

É uma doença com critérios de diagnósticos bem definidos, porém de manejo complexo, uma vez que sua abordagem, além da terapêutica medicamentosa, envolve uma série de mudanças nos hábitos de vida dos pacientes (SANTOS; ROSSI; NASCIMENTO, 2010).

Em muitos países, a prevalência do *diabetes mellitus* tem se elevado vertiginosamente e espera-se ainda um maior incremento. Nos países em desenvolvimento, há uma tendência de aumento dos casos em todas as faixas etárias, especialmente nas mais jovens, cujo impacto negativo sobre a qualidade de vida e a carga da doença aos sistemas de saúde é imensurável (SARTORELI; FRANCO, 2009).

O diabetes se associa a grandes cargas econômicas e sociais, tanto para o indivíduo como para a sociedade. Seus custos estão relacionados principalmente com uma alta frequência de complicações agudas e crônicas, que são causas de

hospitalização, incapacitações, perda de produtividade de vida e morte prematura (HARRIS, 1998). Além de atingir em todo o mundo grande número de pessoas de qualquer condição social, o diabetes representa um problema pessoal e de saúde pública com grandes proporções quanto à magnitude e à transcendência, apesar dos progressos no campo da investigação e da atenção aos pacientes (GIGANTE; ASSUNÇÃO; SANTOS, 2001).

Nas Américas, o número de indivíduos com diabetes foi estimado em 35 milhões no ano 2000 e projetado para 64 milhões em 2025. Nos países desenvolvidos, o aumento ocorrerá principalmente nas faixas etárias mais avançadas, decorrente do aumento da esperança de vida e do crescimento populacional; nos países em desenvolvimento, o aumento será observado em todas as faixas etárias, principalmente no grupo de 45-64 anos onde sua prevalência deverá triplicar, duplicando nas faixas etárias de 20-44 e 65 e mais anos (KING; AUBERT; HERMAN, 1998).

No Brasil, as cidades das regiões Sul e Sudeste, consideradas de maior desenvolvimento econômico do país, apresentam maiores prevalências de *diabetes mellitus* e de tolerância à glicose diminuída. Os principais fatores associados à maior prevalência do diabetes no Brasil foram: a obesidade, o envelhecimento populacional e o histórico familiar de diabetes (MALERBI; FRANCO, 1992).

A mortalidade proporcional por *diabetes mellitus* também tem mostrado um importante crescimento, quando comparada a outras afecções. Há estudos que demonstram que o diabetes como causa de morte tem sido subnotificado, pois os diabéticos geralmente morrem devido às complicações crônicas da doença, sendo estas que figuram como causa do óbito (FRANCO et al., 1998).

Entre os fatores envolvidos na etiologia das complicações crônicas do *diabetes mellitus* destacam-se a hiperglicemia, a hipertensão arterial sistêmica, a dislipidemia e o tabagismo. Além destes, outros fatores de risco não

convencionais têm sido descritos: disfunção endotelial, estado pré-trombótico e inflamação (SCHEFFEL et al., 2004).

O comprometimento aterosclerótico das artérias coronarianas e dos membros inferiores é comum nos pacientes com *diabetes mellitus* e constitui a principal causa de morte destes pacientes. Estas complicações macroangiopáticas podem ocorrer mesmo em estágios precoces do diabetes e se apresentam de forma mais difusa e grave do que em pessoas sem diabetes. Além disso, pacientes com *diabetes mellitus* podem apresentar problemas de visão, doença renal e dano neuronal, que são chamadas de complicações microangiopáticas (SCHEFFEL et al., 2004).

As lesões de extremidades inferiores nos pacientes diabéticos constituem um grande problema de saúde pública, por serem frequentes na população diabética de baixo nível sócioeconômico, com condições inadequadas de higiene e pouco acesso aos serviços de saúde. Quando os pacientes procuram atendimento médico, as lesões geralmente estão em estágios avançados, requerendo tratamento cirúrgico, que muitas vezes os incapacitam para suas atividades de rotina. Lesões corriqueiras evoluem desfavoravelmente, principalmente porque a sensibilidade diminuída nas extremidades, associada à deficiência visual, interfere na percepção de pequenos traumas e feridas (SCHEFFEL et al., 2004).

2.1.1 Pé diabético

Scapim (2004) refere que as úlceras nos pés são erosões cutâneas caracterizadas pela perda do epitélio, que invadem a derme, atingindo os tecidos profundos. São exemplos dessas úlceras as imagens mostradas na Figura 1. Essas úlceras resultam de vários fatores etiológicos e podem agravar-se pela inabilidade de reparação tecidual de maneira oportuna e ordenada. Uma boa

revisão sobre o problema do pé diabético pode ser encontrada em Duarte e Gonçalves (2011).

Figura 1 Imagens do pé diabético

Fonte: Duarte; Gonçalves (2011).

Contaminações acometendo os pés são as principais infecções de partes moles em diabéticos. Além disso, tais contágios são também a principal causa de amputação, acarretando altos custos à sociedade. É importante citar que amputações não traumáticas ocorrem 100 vezes mais frequentemente em pacientes diabéticos (ROCHA et al., 2002 apud SCAPIM, 2004).

Este quadro clínico pode ser evitado, mediante alguns cuidados. O reconhecimento precoce do pé em risco às úlceras e a prontidão do cuidado destas são primordiais para diminuir o impacto da doença. É responsabilidade dos profissionais de saúde realizar esse reconhecimento precoce, porém, nem sempre é cumprido, ocasionando assim amputações em pacientes que não realizaram exames completos ou nunca haviam recebido informações necessárias (FARJADO, 2006).

A preocupação com o pé diabético vem aumentando no Brasil. Especialistas ressaltam a atenção especial com os pacientes diabéticos, que nem sempre a neuropatia diabética manifesta sintomas e um diagnóstico precoce pode ser a chave para evitar danos aos pés. Ainda é enfatizado que o quadro clínico do paciente vai piorando com o tempo: quando há um diagnóstico

precoce, o controle glicêmico pode reverter a situação. Caso não seja tomada alguma atitude, pode ser necessária a amputação de dedos ou dos pés (DISSAT; RODRIGUES, 2013).

2.2 Métodos de agrupamentos

Nesta seção, será apresentado os métodos de agrupamentos utilizados no projeto.

2.2.1 *K-means*

K-means é uma técnica de agrupamento que utiliza medidas de similaridade para criar um particionamento de um conjunto de dados em subconjuntos que apresentem alguma semelhança (TAN; STEINBACH; KUMAR, 2009).

O *K-means* foi utilizado pela primeira vez por James MacQueen em 1967, porém, ele foi primeiramente proposto por Stuart Lloyd em 1957 (LLOYD, 1982) e publicado no ano de 1982 (MACQUENN, 1967).

O método *K-means* toma o parâmetro de entrada k (número de centroides) como o número de agrupamentos (*Clusters*) e partições de um conjunto de dados de n objetos. Em seguida cada objeto é atribuído a um agrupamento com base na proximidade do objeto com o centroide que representa aquele agrupamento. Para atribuir o objeto para o centroide mais próximo, é utilizada uma métrica, como a distância Euclidiana, para calcular a distância dos objetos aos centroides. Depois que todos os objetos são distribuídos para k grupos, os centroides são atualizados verificando a melhor posição e garantindo a menor distância entre seus dados e a maior distância entre os dados dos outros grupos, tomando a média de objetos de k grupos, respectivamente. O processo é repetido até que não haja nenhuma mudança nos

k centroides. *K-means* visa à partição de *n* observações em conjuntos, de modo a minimizar a soma dos quadrados dos erros dentro do conjunto:

$$e= \sum_{i=1}^{K} \sum_{\mathbf{p} \in C_i} |\mathbf{p}-\mathbf{m}_i|^2 , \quad (1)$$

Onde:

e a soma das distâncias euclidianas entre cada objeto e o centro da classe que ele pertence.

p é o objeto que pertence a um grupo C_i e

$\mathbf{m}_i$ é a média do conjunto C_i.

O algoritmo inicia com *k* centroides (que podem ser definidos aleatoriamente ou baseados em alguma informação *a priori*). Então, iterativamente:

a) são calculadas as distâncias de todos os elementos do conjunto de dados em relação aos k centroides, os elementos de menor distância em relação aos centroides formam grupos e;
b) são atualizados os centroides em relação a cada grupo.

O algoritmo *K-means* é formalmente descrito no Algoritmo 1 (TAN; STEINBACH; KUMAR, 2009).

Algoritmo *K-means* Básico

Selecione *k* pontos como centroides iniciais.

Repita

I. Forme *k* grupos atribuindo cada ponto ao centroide mais próximo, a partir de uma métrica de cálculo de distância.

Recalcule o centroide de cada grupo.

Até que

Os centroides não mudem.

Fim

Algoritmo 1 Algoritmo *K-means* básico

Para ilustrar o funcionamento do algoritmo, considere o conjunto de dados apresentado na Figura 2. O objetivo é encontrar três grupos *(clusters)* definidos dentre os ponto, a Figura demonstra a importância de escolher centroides iniciais, e a partir de seis atualizações os grupos são encontrados. Os centroides são representados pelo símbolo (+).

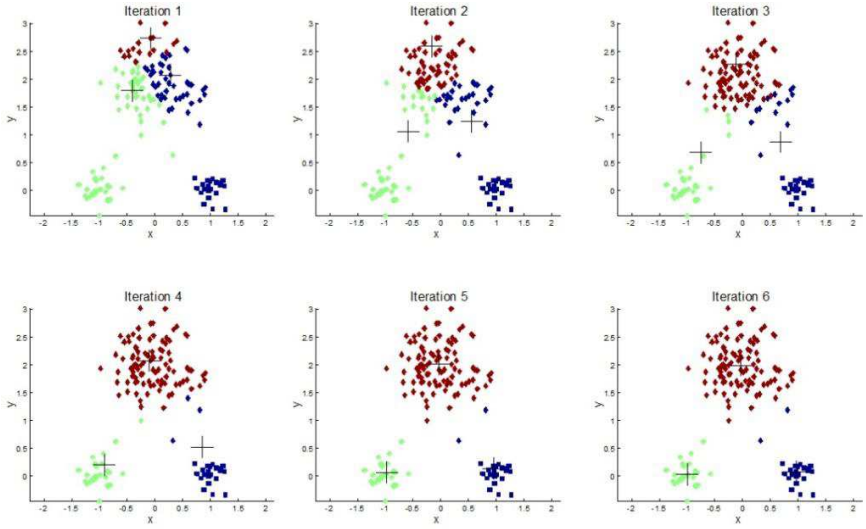


Figura 2 Usando o algoritmo *K-means* para encontrar três grupos nos dados de exemplo

Fonte: Pandre (2014).

Segundo Turi (2001), as principais vantagens deste algoritmo são a fácil implementação e o fato do mesmo se basear na intuição sobre a natureza dos exemplos. O algoritmo *K-means* é sensível ao tamanho dos grupos, a densidade de pontos de dados, formas não-globular de grupos e, claro, a *outliers*, mas em combinação com visualização de dados adequada desses problemas, podem ser resolvidos na maioria dos casos.

2.2.2 Modelo de Mistura de Gaussianas

De acordo com Reynolds (2007), um modelo de Misturas de Gaussianas (*Gaussian Mixture Models*) é uma função de densidade de probabilidade representada pela soma ponderada de componentes de densidades gaussianas. A distribuição gaussiana requer apenas dois parâmetros, a média μ e variância σ^2,

tornando assim a utilização generalizada das misturas gaussianas (ZHUANG et al., 1996).

Uma possível representação dos grupos é matematicamente, e ocorre a partir de distribuição paramétrica, como a Gaussiana (contínua) (DEMPSTER, 1977) ou uma Poisson (discreta) (SASLAW, 1989).

GMDistribution (*Gaussian mixture distribution*) é usado para estimar automaticamente os parâmetros de um modelo de mistura Gaussiana a partir de dados de amostra (FRALEY; RAFTERY, 2002). O *GM Distribution* aplica o algoritmo de *expectation-maximization* (EM), juntamente com uma estratégia de agrupamento aglomerativo para estimar o número de grupos que melhor se ajustam aos dados (CELEUX; GOVAERT, 1995). A estimativa é baseada nos critérios de identificação ordem Rissenen conhecidos como MDL Comprimento Mínimo de Descrição (*Minimum Description Length*).

Modelos de misturas gaussianas são utilizados para agrupamento de dados, eles são formados a partir da combinação de componentes de densidade normais multivariadas. Como o *K-means*, o *GMDistribution* utiliza um algoritmo iterativo que converge até obter o melhor resultado.

2.2.3 Validação de agrupamentos

Nesta seção, serão apresentados os métodos de validação de agrupamentos utilizados no projeto.

2.2.3.1 *Silhouettes*

Uma ferramenta importante na avaliação da qualidade de agrupamentos *(clusters)* em sistemas de reconhecimento de padrões é conhecida por *Silhouettes* (ANTONELLI et al., 2013). Essa ferramenta funciona como um indicador de qualidade entre homogeneidade intra-*cluster* e inter-*cluster* e foi

desenvolvida por Rousseeuw em 1987 (ROUSSEEUW, 1987). Trata-se de um método de interpretação e validação de conjunto de dados através de um gráfico que representa como cada objeto encontra-se dentro de seu grupo. A partir de um conjunto de dados que foi agrupado em *k* grupos, por uma técnica de *clusterização*, para cada ponto de referência *i,* tem-se*:*

$$S_i = \frac{b_i - a_i}{\max(b_i - a_i),} \quad , \tag{2}$$

Em que a_i é a distância média de um paciente (*i)* pertencente a um grupo em relação a todos os outros pacientes dentro do mesmo grupo, e b_i é a menor distância média de seus vizinhos. A equação (2) pode ser descrita como:

$$\begin{cases} 1 - a_i / b_i & \text{se } a_i < b_i, \\ S_i = \quad 0 & \text{se } a_i = b_i, \\ (a_i / b_i) - 1 & \text{se } a_i > b_i. \end{cases} \tag{3}$$

A partir da definição acima, é evidente que:

$$-1 \leq s_i \leq 1$$

Para adquirir um s_i mais próximo de 1, exige-se $a_i < b_i$. Como a_i é uma medida de dissimilaridade de *i* para seu próprio grupo, um valor pequeno significa que ele está bem adaptado. Além disso, um valor alto para b_i implica que *i* está mal associado ao aglomerado vizinho. Assim, um s_i perto de 1 significa que os dados estão devidamente agrupados. Se s_i está perto de um negativo, ou seja, um valor de a_i maior que o valor de b_i, então, pela mesma

lógica, vê-se que *i* seria mais adequado se fosse agrupado no seu conjunto vizinho. Um s_i perto de 0 significa que o dado está na fronteira dos grupos.

2.2.3.2 Dendrograma

As representações gráficas dos resultados de agrupamentos obtidos constituem uma importante etapa da análise de agrupamentos, pois facilita a percepção dos grupos obtidos (FREI, 2006). Dendrograma é uma representação gráfica de diferentes agregações feitas durante o processo de análise de grupos. Consiste de nós que correspondem a grupos de ramos que representam associações feitas a cada passo. A estrutura do dendrograma é determinada pela ordem em que cada agregação é feita (DODGE, 2008).

O dendrograma é bastante utilizado na apresentação de resultados de agrupamentos, pelo fato de ser bidimensional, o que facilita a interpretação, relata Sneath e Sokal (1973). O primeiro exemplo de dendrograma foi utilizado no trabalho de Ernst Mayr em 1953 (MAYR, 1953).

Um dendrograma é composto por muitas linhas em forma de U que conectam os pontos de dados em uma árvore hierárquica. A similaridade entre dois objetos em um dendrograma é representada pela altura do nó interno mais baixo que eles compartilham. A altura de cada linha representa a distância entre os pontos de dados. (IZENMAN, 2008). Na Figura 3, é apresentada a forma básica de um dendrograma com suas respectivas terminologias.

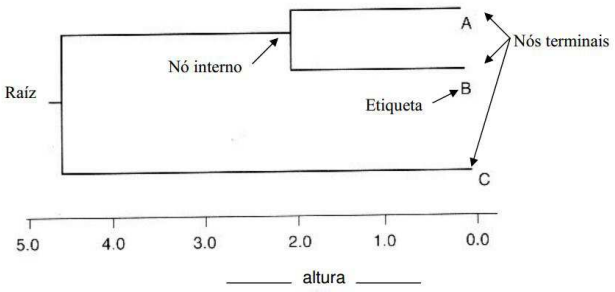


Figura 3 Forma de um dendrograma

Fonte: Quintal (2006).

Os nós internos do dendrograma são formados pela união dos nós terminais.

A Figura 4 exibe um dendrograma contendo 10 dados.

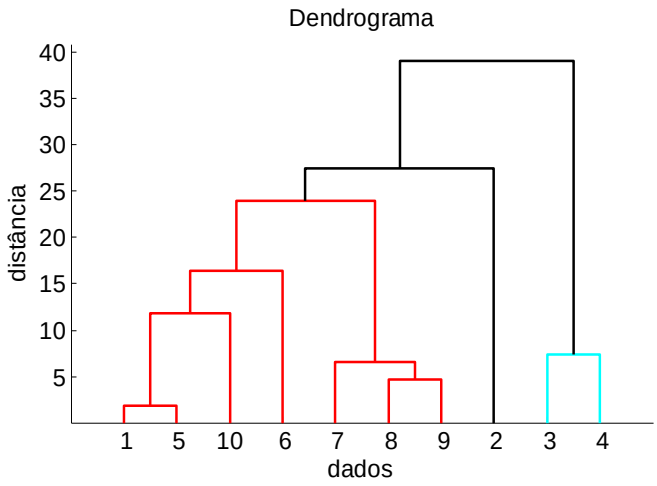


Figura 4 Exemplo de dendrograma

Para o agrupamento destes, usou-se a distância euclidiana, indicada na abscissa. Observa-se que os dados (1, 5) unidos por colchetes são os mais

similares (distância euclidiana igual a 1). Os dados (6, 10), (7, 8) que estão agrupados, unindo com os grupos (1, 5) e 9, formam um grupo maior, e unindo ao 2 finaliza o primeiro grupo, considerando a distância euclidiana entre 25 e 30. Ficando os dados (3, 4) para o segundo grupo, assim por meio desta representação, pode-se vislumbrar que os dados se dividem em 2 grupos (*clusters*).

2.3 Classificadores Supervisionados

Nesta seção, serão apresentados os métodos de classificação supervisionados utilizados no projeto.

2.3.1 K Vizinhos mais próximos (K *Nearest Neighbor*)

O método k-vizinhos mais próximos (KNN, do inglês *k-Nearest Neighbors*) é considerado um dos métodos de classificação mais antigos e simples (COVER; HART, 1967). Apesar da sua simplicidade, esse método tem alcançado bom desempenho em diferentes cenários (BELONGIE; MALIK; PUZICHA, 2002; SIMARD; LECUN; DENKER, 1992).

O algoritmo KNN é um algoritmo de aprendizado supervisionado do tipo *lazy* (aprendiz preguiçoso). Um aprendiz preguiçoso simplesmente armazena os eventos de treino e realiza uma única etapa para classificar eventos (AHAH; KIBLER; ALBERT, 1991). A ideia geral desse algoritmo consiste em encontrar os *k* eventos rotulados mais próximos do exemplo não classificado através do cálculo de uma distância e, com base no rótulo desses eventos mais próximos, é tomada a decisão relativa à classe do evento não rotulado. A proximidade é normalmente definida em termos de distância Euclidiana, no entanto outras distâncias podem ser utilizadas. Para a aplicação da Distância Euclidiana, é necessário definir dois conjuntos que representam os espaços

celulares **P** ($p_1, p_2, ... p_n$) e **Q** ($q_1, q_2, ... q_n$) em que **Q** representa o espaço celular contendo os dados $q_1, q_2, q_3 ... q_n$ e **P** representa o espaço celular contendo os dados $p_1, p_2, p_3, ... p_n$. A partir destes conjuntos, pode-se determinar uma distância d(**Q**, **P**) como sendo:

$$d(\mathbf{P},\mathbf{Q}) = \sqrt{\sum_{i=1}^{n} (\mathbf{P}_i - \mathbf{Q}_i)^2} \quad . \qquad (4)$$

Os algoritmos da família KNN requerem pouco esforço durante a etapa de treinamento. Em contrapartida, o custo computacional para rotular um novo exemplo é relativamente alto, pois, no pior dos casos, esse exemplo deverá ser comparado com todos os exemplos contidos no conjunto de exemplos de treinamento.

Dado um evento de teste *d*, para classificá-lo o método KNN tradicionalmente realiza as seguintes atividades:

Algoritmo *KNN* Básico
Selecione *k* vizinhos.
    Faça
        Calcule a distância entre o evento *d* a cada um dos eventos de treino utilizando uma métrica, tal como a medida da distância Euclidiana.
        Selecione os *k* eventos mais próximos, isto é, os mais similares ao evento *d*.
    Classifique o evento *d* como sendo da mesma classe da maioria dos eventos *k* rotulados mais próximos.
    Fim

Algoritmo 2 Algoritmo *KNN* básico

Três parâmetros importantes devem ser determinados para a execução de KNN:

a) quais eventos rotulados, i.e., eventos de treinamento, devem ser lembrados?
b) qual a medida que quantifica a distância entre o exemplo não classificado e os eventos de treinamento?
c) quantos/quais vizinhos mais próximos devem ser considerados?

Na Figura 5, é ilustrado o procedimento de classificação via KNN, com um conjunto de exemplos de treinamento descrito por dois atributos, no qual, eventos com rótulo positivo (+) referem-se a pacientes doentes e eventos com rótulo negativo (−) a não doentes. Considerando o algoritmo KNN para classificação, com $k = 1$ (um vizinho mais próximo), o novo exemplo *Ei* seria classificado de acordo com o vizinho mais próximo, que é o da classe positiva (+).

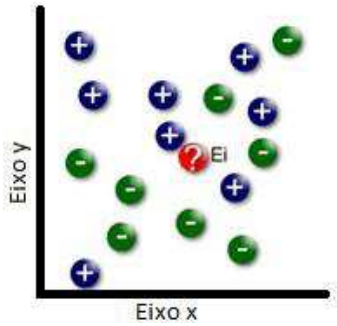
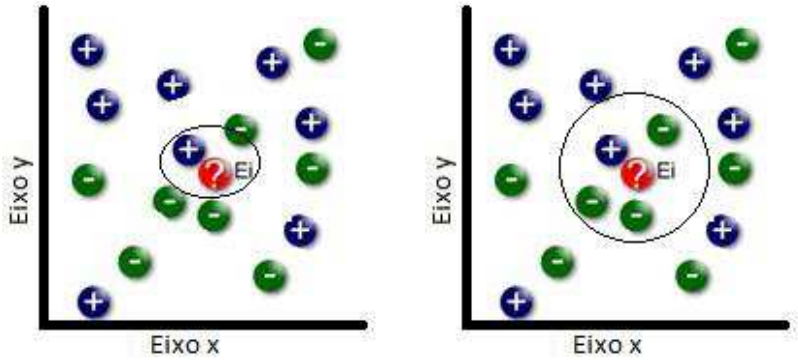


Figura 5 Exemplo de classificação do método *K-Nearest Neighbor*

O algoritmo KNN classifica eventos considerando a classe dos k vizinhos mais próximos. Se $k = 1$, então o exemplo é classificado com a mesma

classe do exemplo mais próximo segundo a medida de distância utilizada. Se $k > 1$, então são consideradas as classes dos k eventos mais próximos para realizar a classificação. Nesse caso, a abordagem mais simples consiste em atribuir ao evento a classe majoritária (predominante) dos k eventos mais próximos.

Na Figura 6, são ilustrados ambos os casos na classificação do evento *Ei*, utilizando um conjunto de eventos positivos (+) e negativos (−) descritos por dois atributos. No primeiro caso (k=1), o exemplo *Ei* será classificado como positivo. Já no segundo caso (k=4), a maioria dos quatro eventos mais próximos é negativa e *Ei* será classificado como negativo.



(a) Exemplo de k=1 (b) Exemplo de k=4

Figura 6 Exemplos de aplicação do algoritmo KNN, com k = 1 e k = 4

Como pode ser observado, o número de vizinhos mais próximos a ser considerado na classificação de novos exemplos influencia fortemente a classificação. É importante ressaltar que não existe um único valor de k que seja apropriado para todos os problemas, de modo que esse valor deve ser avaliado para cada problema em particular. No caso de simplesmente utilizar a classe majoritária dos k eventos mais próximos para classificar eventos, valores

ímpares de *k* são mais apropriados a fim de não ter situações de empate. Outras abordagens alternativas consistem na atribuição de pesos a cada um dos *k* vizinhos mais próximos de acordo com a medida de distância considerada, *i.e.*, os *k* vizinhos mais próximos são ordenados em ordem crescente de similaridade com o evento a ser classificado, tal que, para a determinação da classificação, a classe dos eventos de maior similaridade tem peso maior que a classe dos eventos de menos similaridade.

A quantidade de eventos de treinamento a serem lembrados tem influência direta no tempo de busca pelos eventos mais próximos do evento a ser classificado, pois é necessário comparar esse evento com todos os armazenados (ALPAYDIN, 2004). Dependendo do domínio, essa quantidade de eventos pode ser muito grande e tornar o processo de classificação lento, até o ponto de não atender ao requisito de tempo máximo de resposta para determinado problema.

Ao invés de utilizar muitos eventos de treinamento, o ideal é armazenar somente os eventos mais representativos de cada classe, resumindo a informação mais importante em um conjunto menor de eventos. Em AHAH; KIBLER e ALBERT (1991) são descritas algumas estratégias para selecionar os exemplos mais representativos de cada classe, a partir do conjunto de eventos rotulados disponíveis, contribuindo para a redução do custo para classificar novos eventos e do espaço ocupado em memória pelos eventos de treinamento.

2.3.2 Redes Neurais

As Redes Neurais Artificiais (RNA) (HAYKIN, 2008) são modelos paramétricos não lineares (SANTOS et al., 2003). Imitando o cérebro humano, as RNA são computacionalmente elaboradas. São máquinas projetadas para modelar o processamento do cérebro humano numa função específica. Esta

implementação pode se fazer por meio de "*hardware*" ou simplesmente ser simulada por programação. O grande interesse de diversos pesquisadores pelas RNAs se dá pela sua utilidade e capacidade de aprendizado e generalização. A capacidade da rede em produzir saídas adequadas para entradas que não estavam presentes durante o treinamento, assim sendo possível a resolução dos problemas apresentados computacionalmente complexos (GOULARTE et al., 2006).

As RNAs possuem a incrível capacidade de aprenderem com o meio através de um processo adaptativo conhecido por aprendizagem. De acordo com Haykin (2008), o processo de aprendizagem é realizado por meio de um algoritmo, se assemelhando ao cérebro em dois aspectos:

a) O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem. Isso significa que a rede não atua em um ambiente desconhecido ou desordenado.
b) Forças de conexão entre neurônios, os pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Algumas das propriedades úteis de redes neurais são:

a) Não linearidade
b) Mapeamento de Entrada-Saída
c) Adaptabilidade
d) Tolerância a falhas
e) Uniformidade de análise e projeto

O neurônio, componente principal e fundamental de uma rede neural, como no cérebro humano, pode ser definido sistematicamente conforme ilustra a Figura 7.

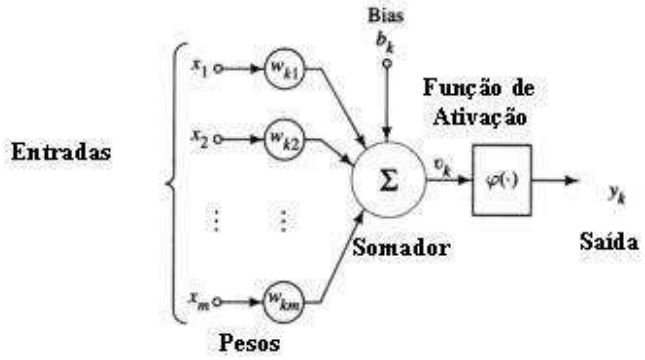


Figura 7 Modelo de Neurônio Artificial
Fonte: O NEURÔNIO... (2014).

Das diversas arquiteturas (estruturas) de redes, a mais utilizada em aplicações de reconhecimento de padrões é a rede alimentada diretamente (*feedforward*) com múltiplas camadas. Conforme definido em Haykin (2008), esta rede consiste de um conjunto de unidades sensoriais que constituem a camada de entrada, uma ou mais camadas ocultas e uma camada de saída. O sinal de entrada se propaga para frente através da rede, camada por camada. Tais redes são, normalmente, chamadas de *perceptrons* de múltiplas camadas (MLP – *Multilayer Perceptron*) (ROSENBLANTT, 1962). A Figura 8 ilustra um *perceptron* com quatro camadas, sendo duas camadas ocultas.

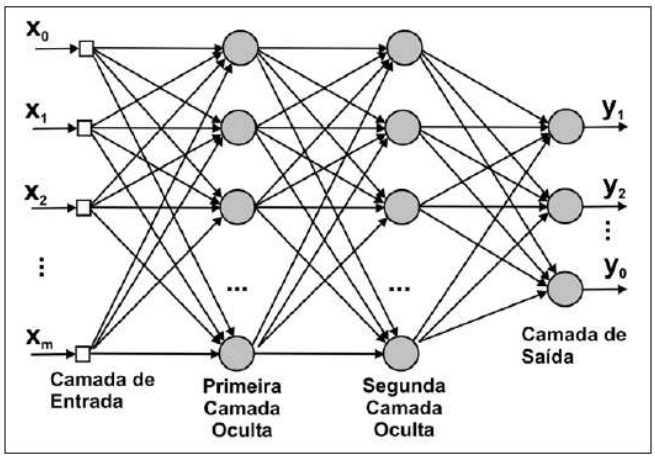


Figura 8 *Perceptron* de múltiplas camadas
Fonte: O NEURÔNIO... (2014)

Os *perceptrons* de múltiplas camadas têm sido aplicados com bastante sucesso para resolver diversos problemas complexos, através do seu treinamento de forma supervisionada com um algoritmo muito popular conhecido como algoritmo de retropropagação de erro (*error back-propagation*) (HAYKIN, 2008).

As RNAs extraem informações relevantes de padrões de informações que lhe forem apresentadas, criando assim uma representação própria. Esta etapa é conhecida por aprendizagem ou treinamento, e consiste em um processo iterativo de ajuste de parâmetros da rede, dos pesos de conexões entre as unidades de processamento que guardam, ao final do processo, o conhecimento que a rede adquiriu do ambiente que está operando.

Nesta etapa, um algoritmo bastante empregado é o algoritmo *Resilient Propagation* (Rprop) (RIEDMILLER; BRAUN, 1993), que tem capacidade de acelerar o processo de aprendizagem. A principal característica deste algoritmo é que os ajustes dos pesos (ω) e da taxa de aprendizado (η) dependem apenas dos

sinais dos gradientes da função erro $E(\omega)$, não dependendo, portanto, da sua amplitude. A função E (ω) é responsável pela especificação de um critério de desempenho que está associado à rede (HAYKIN, 2008).

No algoritmo Rprop, os pesos e a taxa de aprendizagem são alterados apenas uma única vez a cada época de treinamento. Cada peso (ω_{ij}) possui sua própria taxa de variação Δ_{ij}, a qual varia conforme a equação seguinte (RIEDMILLER; BRAUN, 1993):

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+}\Delta_{ij}^{(t-1)}, & \text{se} \quad \frac{\delta E}{\delta \Delta w_{ij}}^{(t-1)} \frac{\delta E}{\delta \Delta w_{ij}}^{(t)} > 0 \\ \eta^{-}\Delta_{ij}^{(t-1)}, & \text{se} \quad \frac{\delta E}{\delta \Delta w_{ij}}^{(t-1)} \frac{\delta E}{\delta \Delta w_{ij}}^{(t)} < 0 \\ \Delta_{ij}^{(t-1)}, & \text{caso contrário} \end{cases} \tag{5}$$

Em que, $0<\eta^{-}<1<\eta^{+}$. Assim, cada vez que a derivada parcial do peso correspondente ω_{ij} alterar o sinal significa que a última atualização foi muito excessiva e o algoritmo pulou um mínimo local. Como consequência, o valor de Δ_{ij} é reduzido pelo fator η^{-}. Se o sinal da derivada se mantém, o valor de Δ_{ij} é levemente aumentado para acelerar o processo de convergência. Os pesos da rede são então alterados da seguinte forma:

$$\omega_{ij}^{(t+1)} = \omega_{ij}^{(t)} + \omega_{ij}^{(t)} \tag{6}$$

Em que $\omega_{ij}^{(t)}$ é definido por,

$$\omega_{ij}^{(t)} = \begin{cases} -\omega_{ij}^{(t)} \text{ , se } \frac{\delta E}{\delta \Delta w_{ij}}^{(t)} > 0 \\ +\omega_{ij}^{(t)} \text{ , se } \frac{\delta E}{\delta \Delta w_{ij}}^{(t)} < 0 \\ 0 \text{ , caso contrário.} \end{cases} \quad (7)$$

Na área da saúde, as redes neurais artificiais representam um paradigma metodológico muito utilizado. Sua aplicação vem se tornando extremamente eficiente e eficaz em inúmeras áreas da medicina, principalmente na área de diagnóstico, prognóstico e terapia. Devido ao fato de não haver necessidade de independência e normalidade das variáveis em estudo, bem como a grande capacidade de aprendizado a partir do ambiente, a aplicação de redes neurais artificiais na análise estatística de dados epidemiológicos tem tido grande aceitação. Além do mais, o processamento neural é capaz de extrair correlações das variáveis de entrada diretamente sobre os espaços de dimensão elevada que tipicamente as caracterizam, tornando tal processamento uma ferramenta valiosa em problemas complexos de reconhecimento de padrões (SANTOS et al., 2005).

2.3.3 Árvore de Decisão

As Árvores de Decisão são ferramentas comumente utilizadas para dar a um algoritmo de aprendizado a capacidade de aprender e de tomar decisões, gerando um gráfico em forma de árvore. Cada ramo da árvore de decisão representa uma possível decisão ou ocorrência (HAN; KAMBER, 2005).

Também conhecidas como árvores de classificação, as AD foram popularizadas na comunidade estatística através do trabalho de Breiman et al. (1984). Eles propuseram um modelo de decisão por árvore binária, conhecido como CART (*Classification and Regression Tree*) (DUDA; HART; STORK, 2000). Tal modelo descreve a distribuição condicional da variável resposta dada,

onde as variáveis respostas assumem valores categóricos. As variáveis explicativas assumem valores contínuos ou categóricos (SANTOS et al., 2003).

A árvore de decisão é considerada uma técnica de mineração de dados, cuja motivação consiste em descobrir conhecimento a partir de uma base de dados Pode ser utilizada para classificação de dados, predição de saídas e geração de regras de classificação de fácil compreensão. Além disso, elas também possibilitam a visualização gráfica das consequências das decisões (SANTANA, 2005). Elas podem ser utilizadas nas mais diversas áreas do conhecimento por apresentarem poucas restrições quanto às características das variáveis adotadas, de tal modo que não exigem distribuição normal, além de admitirem a dependência entre as variáveis (LIN; CHEN, 2011).

As AD são capazes de converter o conhecimento em regras; para isso é necessário que seja feito o particionamento ou classificação da AD, objetivando responder, inicialmente, qual dos atributos (variáveis constantes na base de dados) será o nó raiz (SILBERSCHATZ et al., 2006).

A função precípua da AD consiste em particionar recursivamente um conjunto de treinamento, de modo que cada subconjunto obtido apresente casos de uma única classe. Diz-se, portanto, que utiliza o paradigma “dividir para conquistar”, pois o problema principal é dividido em subproblemas até que a solução seja encontrada (CASTANHEIRA, 2008).

Segundo Tan, Steinbach e Kumar (2005), uma árvore de decisão possui três tipos de nodo:

a) um nodo raiz, que não possui nenhuma aresta de entrada e zero ou mais arestas de saída;
b) nodos internos, cada qual com exatamente uma aresta de entrada e duas ou mais arestas de saída;

c) nodos folha ou terminal, cada qual com uma única aresta de entrada e nenhuma de saída, pois é o nodo que determina a qual classe o exemplo pertence.

Na Figura 9, é ilustrado um exemplo de árvore de decisão para diagnosticar pacientes como doente ou saudável.

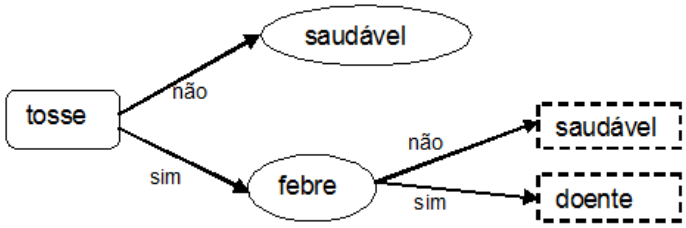


Figura 9 Exemplo de Árvore de Decisão

Para classificar o paciente utilizando uma árvore de decisão, basta começar pelo nodo raiz da árvore (tosse) em que é aplicado o primeiro teste com o atributo referente a este nodo, o paciente está ou não com tosse. O processo se repete até ser encontrado um nodo folha (saudável ou febre) e por fim o último nodo folha (saudável ou doente).

2.4 Algoritmos Genéticos

Os Algoritmos Genéticos (AG), inspirados na natureza, foram criados no final da década de 60, por John Holland. Estudando a evolução das espécies, Holland propôs o modelo heurístico computacional, que depois de implementado poderia oferecer soluções para problemas difíceis que até então eram insolúveis computacionalmente naquela época. Após os trabalhos de Holland e sua influência para os AGs, estes algoritmos começaram a expandir

por toda a área científica, gerando aplicações para soluções de diversos problemas (LINDEN, 2008).

Na área de Inteligência Artificial, tem-se uma subárea chamada Computação Evolucionária, onde diversos "Algoritmos Evolucionários" estão presentes, dentre eles os Algoritmos Genéticos. Eles são baseados na teoria da evolução da espécie proposto por Charles Darwin em 1859 no famoso livro "A Origem das Espécies", teoria essa que introduziu o conceito de seleção natural (OBITKO, 1998).

De acordo com Linden (2008), os "Algoritmos Evolucionários usam modelos computacionais dos processos naturais de evolução como uma ferramenta para resolver problemas" (LINDEN, 2008, p. 37).

O Algoritmo 3 mostra em detalhes os procedimentos para executar um AG:

```
Procedimento AG
{ t = 0;
        inicia_população (P, t)
        avaliação (P, t);
        repita até (t = d)
                { t = t +1;
                seleção_dos_pais (P,t);
                recombinação (P, t);
                mutação (P, t);
                avaliação (P, t);
                sobrevivem (P, t)
        }
}
```

Algoritmo 3 Procedimentos do Algoritmo Genético

onde: *t* - geração atual; *d* - tempo determinado para finalizar o algoritmo; *P* – população.

Apesar da grande variedade dos modelos computacionais, todos têm em comum a simulação da evolução das espécies através da seleção, mutação e

reprodução, sendo chamados de processos. Estes processos dependem do desempenho dos indivíduos dentro do ambiente (LINDEN, 2008).

Em um Algoritmo Genético, considera-se inicialmente uma população de indivíduos (população inicial), sendo que cada indivíduo representa uma possível solução para um problema. A cada iteração, chamada de (geração), esses indivíduos são avaliados em relação a sua adaptabilidade ao meio externo, sendo selecionados os mais aptos para aplicação dos operadores genéticos de cruzamento e mutação. Após aplicação de tais operadores, uma nova população é gerada. O processo é realizado até um individuo "ótimo" ser encontrado, indicando ser a solução do problema (OBITKO, 1998).

Segundo Zuben (2004), um algoritmo genético deve ter basicamente os seguintes componentes:

a) uma representação genética para possíveis soluções do problema;
b) possibilidade de criação da população inicial com as soluções candidatas;
c) uma função de avaliação que classifique as soluções em termos de sua adaptação ao ambiente;
d) operadores genéticos (cruzamento, mutação etc);
e) valores para os diversos parâmetros utilizados pelo algoritmo.

O princípio básico do funcionamento dos AGs é o critério de seleção, esse critério promove indivíduos mais aptos após gerações. Um método de seleção muito utilizado, o Método da Roleta, no qual indivíduos de uma geração são escolhidos para fazer parte da próxima geração, através de um sorteio de roleta (ZUBEN, 2004). Neste método, cada indivíduo da população é representado na roleta proporcionalmente ao seu índice de aptidão. Assim, as maiores porções da roleta pertencem a indivíduos com alta aptidão, e as porções

menores para indivíduos com menos aptidão. O objetivo é girar a roleta um determinado número de vezes, de acordo com o número de indivíduos da população. Aos indivíduos sorteados, operadores genéticos serão aplicados.

Para gerar populações melhores, é necessário um conjunto de operações. Essas operações são o cruzamento (*crossover)* e a mutação. Eles são utilizados para garantir que as novas populações possuam características dos pais, mas que sejam totalmente novas, ou seja, a população se diversifica e mantém características de adaptações anteriores. Para não perder os melhores indivíduos, utiliza-se a reprodução elitista, que automaticamente aloca o melhor indivíduo da geração atual na próxima geração. Esse ciclo é repetido um determinado número de vezes.

Neste projeto os algoritmos Genéticos são utilizados para selecionar variáveis de entrada. O Algoritmo de Seleção visa minimizar os dados de entrada dos algoritmos de classificação de acordo com o necessário para o mesmo alcançar o resultado ótimo. Para isso, o mesmo faz um treinamento selecionando aleatoriamente diversas entradas e treinando o algoritmo de classificação, sempre salvando o melhor resultado. Cada algoritmo possui o número de entradas diferente de acordo com a seleção realizada.

2.5 Reconhecimento de padrões na detecção de doenças

Com relação aos trabalhos que utilizam técnicas de inteligência computacional e reconhecimento de padrões na detecção de doenças, destacam-se, no Brasil, os trabalhos de Santos et al. (2005) e Souza et al. (2007). Santos et al. (2005) propuseram um sistema para predição da soroprevalência da hepatite A utilizando modelos de regressão logística e redes neurais artificiais. Os resultados mostram que o modelo neural, aplicado sobre a informação relevante

extraída do modelo de regressão logística, apresenta um bom desempenho, alcançando uma eficiência de classificação geral acima de 88%.

Souza et al. (2007) utilizou processamento neural para auxiliar o diagnóstico médico na detecção de tuberculose pulmonar. Este sistema, com base num questionário de sintomas, identifica qual é a chance do paciente ter contraído tuberculose, assim como o classifica em um dentre três grupos de risco. Resultados expressivos são obtidos, atingindo-se uma identificação de pacientes doentes de 100%; e de não doentes de 80%.

Na área de *diabetes mellitus*, destacam-se os seguintes trabalhos:

Patil, Joshi e Toshniwal (2010) propuseram um modelo híbrido de predição, utilizando o algoritmos *K-means* (MACQUENN, 1967), a fim de validar rótulos, e posteriormente aplica o algoritmo de classificação C4.5 (QUINLAN, 1993) (algoritmo utilizado para gerar uma árvore de decisão), que constrói um modelo final para classificar a base de dados dos índios Pima, referentes a índios do Arizona, EUA (EKOE et al., 2008), onde mais da metade da população adulta são portadores de diabetes. O objetivo do trabalho foi investigar como os incidentes de diabetes são afetados pelas características e medidas dos pacientes.

Lee et al. (2010) desenvolveram um Sistema de Monitoramento e Assessoria de Gestão em Pacientes Diabéticos utilizando um método baseado em regras e no algoritmo *K-Nearest Neighbor* (KNN) (COVER; HART, 1967). Esse sistema fornece um tratamento adequado para os pacientes diabéticos, de acordo com seu nível de açúcar no sangue.

Karegowda, Jayaram e Manjunath (2012) apresentam um modelo híbrido que classifica o banco de dados de diabetes dos índios Pima. Esse trabalho utiliza o algoritmo *K-means* para identificar e eliminar casos de pacientes classificados incorretamente (DUDA, 2000). Em seguida, Algoritmos

Genéticos (HOLLAND, 1975) são empregados para selecionar variáveis e, por fim, o KNN é utilizado (LINDEN, 2008).

Antonelli et al. (2013) apresentam uma análise de dados que identifica o diabetes através do histórico de exames realizados pelos pacientes. É utilizado o DBSCAN (*Density-based spatial clustering of applications with noise*), um algoritmo baseado em densidade. Esse estudo utilizou a base de dados de pacientes diabéticos fornecidos pelo Centro Nacional de Saúde (CNS) da província de Asti, Itália.

Boyko et al. (2006) propuseram um modelo de previsão de ocorrência de úlcera nos pés utilizando informação clínica comumente disponível, alcançando uma previsão de úlcera do pé diabético em um período de 3 anos com um alto grau de precisão.

3 METODOLOGIA

3.1 Base de Dados

Realizou-se neste trabalho um levantamento de dados de pacientes diabéticos na cidade de Lavras, Minas Gerais, Brasil. Esses dados foram obtidos por meio de um formulário contendo 31 perguntas a respeito do âmbito social, hábitos e principalmente sobre os cuidados com os pés dos pacientes (ANEXO A). Este formulário foi previamente aprovado pelo Comitê de Ética (nº 51.950). As informações foram coletadas durante as visitas domiciliares por um profissional de enfermagem devidamente treinado para tal e um representante do PSF (Posto de Saúde Familiar) no período de outubro a dezembro de 2012. Foram entrevistados um total de 153 pacientes portadores de diabetes com 18 anos ou mais, escolhidos aleatoriamente de quatro unidades básicas de saúde.

3.2 Método proposto

Nesta seção, serão apresentados os métodos propostos no projeto.

3.2.1 Projeto

O projeto do método proposto pode ser dividido em seis etapas, conforme ilustra o diagrama em blocos da Figura 10.

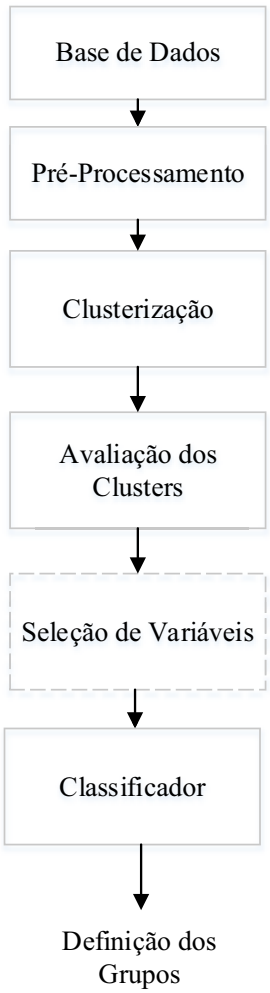


Figura 10 Método proposto: fase de projeto

A primeira etapa é a coleta das informações e formação do banco de dados. As informações foram codificadas de forma que as variáveis assumissem valores entre -1 e 1, seguindo um padrão lógico, desse modo as respostas às perguntas do formulário que apresentaram fatores de risco foram codificados como 1, e as outras respostas como -1. Para questões com mais de duas respostas

possíveis, foram atribuídos valores entre -1 e 1. As questões cujas respostas são unidades de tempo não foram codificadas.

A segunda etapa, de pré-processamento, é responsável por normalizar as variáveis de entrada referentes à idade dos pacientes, tempo (em anos) em que o paciente é diabético e tempo (em anos) em que o paciente trata o diabetes. Essa normalização foi realizada de acordo com (8).

$$\mathbf{z}_i = \frac{\mathbf{x}_i - \mu_i}{\max(\mathbf{x}_i) - \mu_i} \tag{8}$$

Em que $\mathbf{x}_i$ e μ_i são, respectivamente, o vetor da variável *i* e sua média.

Ainda na etapa de pré-processamento, a correlação linear entre as 31 variáveis de entrada foi aplicada com o objetivo de identificar redundância entre as mesmas.

Após a etapa de pré-processamento, algoritmos de agrupamentos de dados são aplicados a fim de encontrar grupos de pacientes com alguma similaridade no espaço de dimensão das informações (variáveis) coletadas. Para este fim, foram utilizados dois algoritmos de agrupamento, o *K-means* (TAN; STEINBACH; KUMAR, 2009) e o Mistura de Gaussianas (FRALEY; RAFTERY, 2002) por serem algoritmos bastante utilizados na literatura e que apresentam, em geral, bons resultados.

A próxima etapa do projeto do método proposto consiste em avaliar os grupos obtidos. Os objetivos desta etapa são: (i) verificar se os grupos encontrados são bem definidos, ou seja, se os pacientes pertencentes a cada grupo estão fortemente agrupados. Para isso, medidas intra e extra grupos são empregadas utilizando-se o método conhecido por *Silhouettes* (ANTONELLI et al., 2013) e o método do Dendrograma (DODGE, 2008); (ii) verificar qual é o número “ideal” de agrupamentos existentes no banco de dados.

Após a avaliação dos grupos obtidos, classificadores são projetados para, automaticamente, classificar um paciente desconhecido como pertencente a um dos grupos obtidos. Para isso, os classificadores baseados nos centroides obtidos pelo *K-means*, redes neurais, árvores de decisão e no algoritmo KNN são empregados, comparativamente.

A fim de reduzir a dimensão dos dados e também aperfeiçoar o desempenho dos classificadores, algoritmos genéticos foram utilizados para selecionar as variáveis mais relevantes, ou seja, aquelas que são capazes de maximizar o desempenho dos classificadores. Essa etapa é importante para revelar quais são as variáveis mais importantes para a identificação do risco de desenvolvimento do pé diabético de acordo com o classificador utilizado.

A fim de validar o método proposto, informações de 30 pacientes fictícios foram simuladas. Buscou-se garantir que 15 pacientes tivessem alto risco de desenvolver pé diabético e 15 pacientes tivessem baixo risco. Para isso, os pacientes fictícios de alto risco foram gerados com as respostas às perguntas do formulário a favor do desenvolvimento do pé diabético (fatores de risco presentes) e os pacientes fictícios de baixo risco foram gerados com as respostas às perguntas, sem a presença de fatores de risco.

Como forma de comparar estes resultados com a opinião de um especialista em diabetes, 20 pacientes foram escolhidos aleatoriamente de cada grupo (alto e baixo risco) e analisados pelo especialista.

3.2.2 Fase operacional

A fase operacional do método proposto é ilustrada na Figura 11. Observe que as etapas de agrupamento, análise de grupo e seleção de variáveis não são mais necessárias. O vetor de informações do paciente é normalizado e apresentado ao classificador.

O resultado da classificação enquadra o paciente em um dos dois grupos de risco de desenvolver o pé diabético: alto ou baixo risco. Se o paciente é classificado como baixo risco, um novo paciente é avaliado. Se o paciente é classificado como alto risco, os devidos cuidados e acompanhamentos são realizados pela equipe de saúde local para evitar o aparecimento do pé diabético.

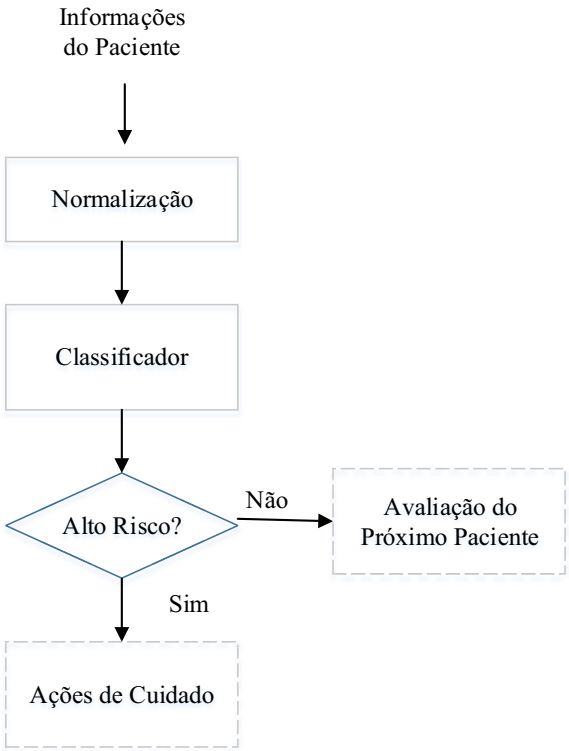


Figura 11 Fase operacional do método proposto

4 RESULTADOS

4.1 Características da População

A Tabela 1 apresenta a caracterização da população de acordo com as informações coletadas. Pode-se notar que a população é composta por pacientes idosos e com baixo nível de escolaridade, o que torna a tarefa de autocuidado mais difícil e, portanto, o trabalho de atenção primária é fundamental para evitar complicações diabéticas.

Tabela 1 Características da População

Perguntas	Estatística
Sexo	68,5% do sexo feminino
Idade (anos)	64,6 ± 13,8
Estado civil	62,5% casados
Ocupação	69,0% aposentados
Educação	95,0% analfabetos ou <8 anos de estudo
Tempo de diagnóstico de diabetes (anos)	11,7 ± 10,5
Diabetes Tipo 1 ou 2	Tipo 2 (91,5%)
Tempo de tratamento da diabetes (anos)	11,6 ± 9,7
Tipo de tratamento	A insulina (5,9%) e a medicação oral (30,7%)
Hábito de examinar os pés	67,5%
Como corta as unhas dos dedos dos pés	Para os cantos e curto demais (47,5%)
O que usa para lavar os pés	Sabonete comum (97,4%)
Hábito de lavar os pés	Diariamente (83,0%)
O que usa para limpar os pés	Toalha comum (83,6%)

"Tabela 1, conclusão"

Perguntas	Estatística
Limpa entre os dedos	76,0%
Usa creme hidratante nos pés	42,0%
O que usa para remover calos	Química (16,4%)
Utiliza bolsa de água quente nos pés	34,0%
Remove cutículas dos dedos dos pés	45,0%
Verifica o calçado dentro antes de usar	65,0%
Tipo de sapato usado	Sapato aberto (65,0%)
A mudança de atitude quando percebe alguma alteração nos pés	65,0% procuram médico ou enfermeiro
Aspecto interno do calçado	Sem costura (52,5%)
Hora do dia para comprar sapatos novos	A qualquer momento (84,3%)
Normalmente caminha com os pés descalços	75,0% nunca descalço
Tipo de meias que utiliza (material)	Algodão (86,3%)
Tipo de meias que utiliza (aparência e costura)	Sem costura e claro (37,9%)
Posição para assistir televisão	Deitada (45,0%)
Material do sapato que usa	Couro sintético (40,0%)
Já percebeu mudanças nos pés	30,0%
Já foi diagnosticado com pé diabético	7,5%

Os dados representam média ± desvio padrão, ou percentual.

4.1.1 Análise de redundância

Como forma de verificar se há redundância entre as 31 variáveis de entrada, foi realizada a correlação linear entre elas. Observou-se que a variável referente ao tempo (em anos) em que o paciente foi diagnosticado como paciente diabético (variável 6), apresentou uma correlação de 0,95 com a variável referente ao tempo (em anos) em que o paciente trata o problema de diabetes (variável 8). A Figura 12 mostra a correlação da variável seis com as demais. Dessa forma, eliminou-se a variável referente ao tempo em que o paciente trata o

problema de diabetes, e passou-se a trabalhar com 30 variáveis. As outras correlações foram inferiores a 0,40 em todos os outros possíveis pareamentos.

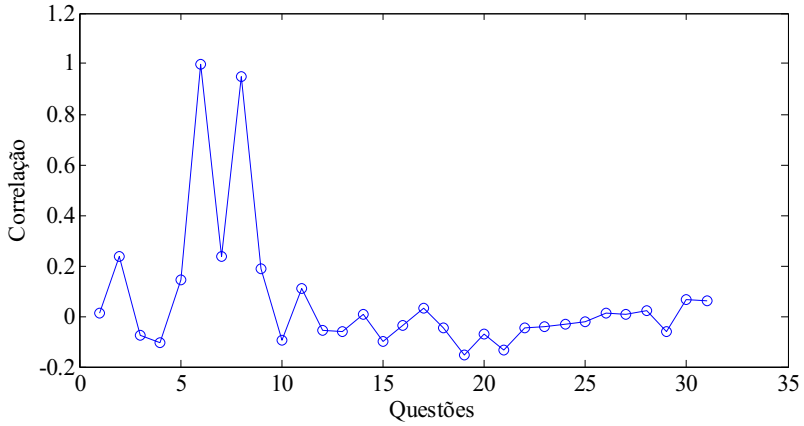


Figura 12 Correlação da variável referente ao tempo (em anos) em que o paciente foi diagnosticado como paciente diabético (identificada como 6) com todas as outras

4.2 Agrupamento

Após a etapa de pré-processamento, os dados foram agrupados em dois e três grupos (*clusters*). O *k-means* foi executado 1.000 vezes, a fim de garantir os melhores *clusters*. A divisão dos pacientes nos grupos pode ser visualizada na Tabela 2.

Tabela 2 Agrupamento dos Pacientes

K-means		***Mistura de Gaussianas***	
Divisão para 2 grupos		**Divisão para 2 grupos**	
Cluster 1	43	*Cluster* 1	54
Cluster 2	110	*Cluster* 2	99
Divisão para 3 grupos		**Divisão para 3 grupos**	
Cluster 1	41	*Cluster* 1	46
Cluster 2	48	*Cluster* 2	55
Cluster 3	64	*Cluster* 3	52

Na divisão em dois grupos, os algoritmos *k-means* e Mistura de Gaussianas obtiveram uma coincidência de 59% dos dados agrupados. Já na divisão em três grupos, somente 33% dos dados foram semelhantemente agrupados.

4.3 Avaliação dos agrupamentos

Como forma de avaliar os agrupamentos obtidos pelos métodos, foi utilizado o *silhouettes*. Os valores de *silhouettes* obtidos para os agrupamentos construídos pelo *k-means* e pela *Mistura de Gaussianas* são mostrados nas Figuras 13 e 14, respectivamente. Valores negativos de *silhouettes* representam colocações inconsistentes de pacientes, enquanto que os valores positivos representam atribuições melhores ao paciente. Observa-se claramente que o *k-means* levou a agrupamentos mais consistentes do que o Mistura de Gaussianas.

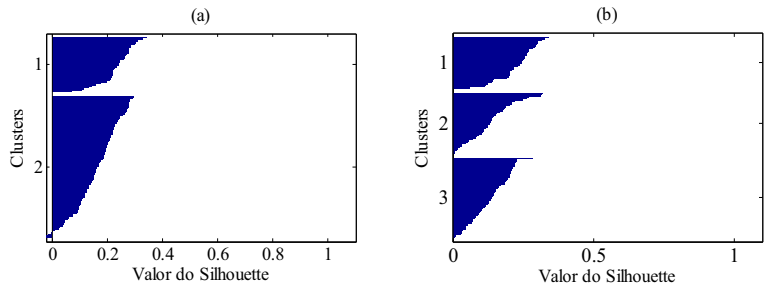


Figura 13 *Silhouettes* para o agrupamento do *K-means,* considerando-se dois agrupamentos (a) e três agrupamentos (b)

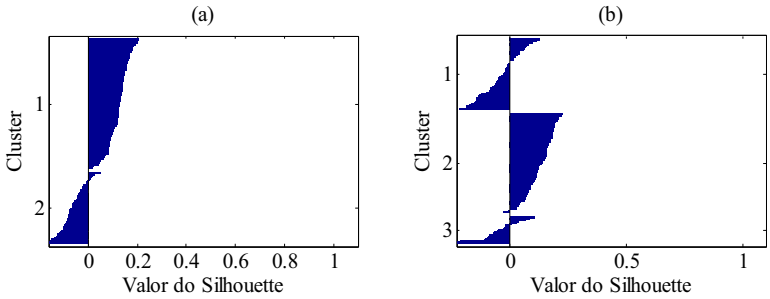


Figura 14 *Silhouettes* para o agrupamento do *GM Distribution*, considerando-se dois agrupamento (a) e três agrupamentos (b)

O cálculo da média dos valores de *silhouettes* para diferentes números de agrupamentos é uma forma quantitativa de avaliar os resultados de *silhouettes* e se ter uma ideia do número de agrupamentos mais adequado para um conjunto de dados. A Tabela 3 mostra essa análise para k = 2, 3, 4 e 5 grupos. Observa-se que o valor de *silhouettes* diminui com o aumento do número de grupos. O gráfico de *silhouettes* para k=4 e 5 não é apresentado por simplificação. A partir destes resultados, apenas o agrupamento do *k-means* e as abordagens para k=2 e 3 foram consideradas.

Tabela 3 Escolha do número de grupos

Análise K-means	
Número de *Clusters k*	**Média das *Silhouettes***
2	0,18 ± 0,01
3	0,16 ± 0,01
4	0,14 ± 0,01
5	0,12 ± 0,02
Análise Mistura de Gaussianas	
Número de *Clusters k*	**Média das *Silhouettes***
2	0,05 ± 0,01
3	0.02 ± 0,01
4	0.00 ± 0,02
5	0.00 ± 0,02

A Figura 15 mostra o resultado do método *silhouettes* quando aplicado ao conjunto de dados referentes aos 30 pacientes fictícios. Observa-se que o grupo dois, referente aos pacientes de baixo risco apresenta um agrupamento levemente mais forte do que o grupo um, com valores de *silhouettes* acima de 0.8.

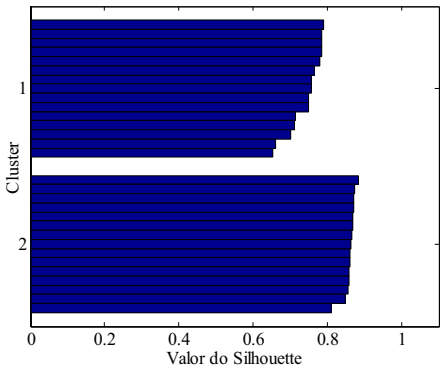


Figura 15 *Silhouettes* para o agrupamento dos 30 dados simulados

Aplicando-se o método do dendrograma aos dados dos pacientes fictícios (Figura 16), observa-se que a divisão foi exata, em dois grupos, separando os 15 pacientes fictícios de alto risco dos de baixo risco.

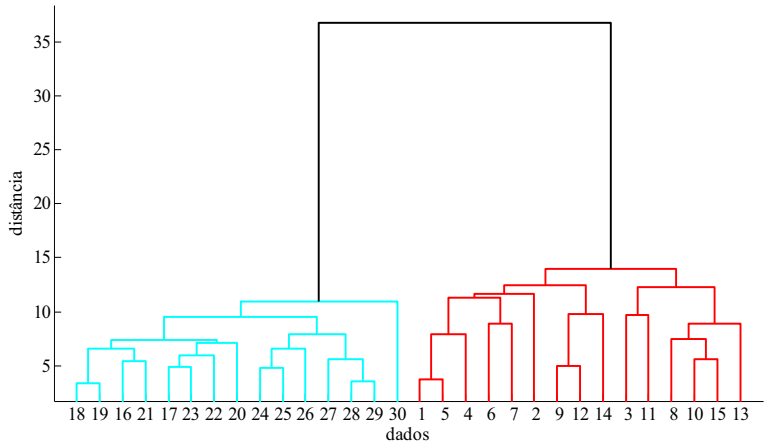


Figura 16 Dendrograma para os 30 dados simulados

Os resultados obtidos pelo método *silhouettes* e dendrograma para os dados de pacientes fictícios são bastante conclusivos, de forma que comprovam a existência de dois grupos de risco bem definidos, alto e baixo risco de desenvolver o pé diabético.

4.3.1 Classificação para dois grupos

A Tabela 4 mostra os resultados de classificação dos 30 pacientes simulados, obtidos pelos classificadores centroides, KNN, AD e RNA considerando-se os dois agrupamentos obtidos pelo *K-means*.

Para o classificador KNN, adotou-se k=5 (vizinhos mais próximos), a distância euclidiana como padrão (não houve diferença nos resultados para

outras métricas de distância), além da regra "aleatória", que é a regra que decide como a amostra será classificada.

Para o classificador RNA, foi utilizado um *perceptron* com topologia 30 x 1. A tangente hiperbólica foi utilizada como função de ativação, para o treinamento foi empregado o algoritmo RPROP *(Resilient Propagation)* (RIEDMILLER; BRAUN, 1993), e por fim 15 épocas foram utilizadas.

Pode-se observar que os classificadores utilizados não foram capazes de separar o *cluster* de pacientes com alto risco de desenvolver o pé diabético do *cluster* de pacientes de baixo risco. O classificador KNN alcançou o melhor desempenho, apresentando uma acurácia de 0,67 ± 0,16.

Tabela 4 Classificação para dois grupos

Classificador	**Nível de risco**	**Número de pacientes**		**Acurácia**
		***Cluster* 1**	***Cluster* 2**	
Centroides	Alto risco	9	6	0,57 ± 0,17
	Baixo risco	7	8	
KNN	Alto risco	11	4	0,67 ± 0,16
	Baixo risco	6	9	
AD	Alto risco	6	9	0,57±0,17
	Baixo risco	4	11	
RNA	Alto risco	9	6	0,53±0,17
	Baixo risco	8	7	

4.3.2 Classificação para três grupos

A Tabela 5 mostra os resultados de classificação dos 30 pacientes simulados, obtidos pelos classificadores centroides, KNN, AD e RNA (*perceptron* multicamadas com topologia 30 x 5 x 3), considerando-se os três agrupamentos obtidos pelo *K-means*. Os classificadores apresentaram bom

desempenho, obtendo uma melhor divisão entre os 30 dados simulados que a abordagem em que apenas dois grupos são considerados.

Tabela 5 Classificação para três grupos

Classificador	Nível de risco	Número de pacientes			Acurácia
		Cluster 1	*Cluster* 2	*Cluster* 3	
Centroides	Alto risco	14	0	1	0,97 ± 0,06
	Baixo risco	0	7	8	
KNN	Alto risco	13	0	2	0,93 ± 0,09
	Baixo risco	0	7	8	
AD	Alto risco	10	0	5	0,78±0,14
	Baixo risco	0	7	8	
RNA	Alto risco	15	0	0	1,00±0,00
	Baixo risco	0	8	7	

O algoritmo KNN usando três agrupamentos (*clusters*) apresentou bom desempenho, em que 13 dos 15 pacientes de alto risco foram classificados como pertencentes ao *cluster* 1 por KNN (Precisão de 0,93 ± 0,09). Já o centroide apresentou melhor desempenho, em que 14 dos 15 pacientes de alto risco foram classificados como pertencentes ao *cluster* 1, e uma precisão de 0,97 ± 0,06. E por fim, a RNA alcançou o melhor resultado, classificando os dados de Alto Risco em um único grupo, e os de Baixo Risco nos outros dois grupos.

Computacionalmente, o uso de centroides leva a um classificador mais simples porque só requer o cálculo de três distâncias para classificar um paciente.

4.3.3 Seleção de variáveis

O AG foi utilizado para selecionar dados de entrada a fim de melhorar a eficiência dos classificadores e apontar as variáveis mais relevantes à

classificação dos pacientes diabéticos como alto ou baixo risco de desenvolver o pé diabético.

Adotou-se como tamanho da população (50), tamanho do indivíduo (30 variáveis de entrada), o número de gerações (100). Foi gerada a função objetivo que é usada para resumir, como uma única figura de mérito como fechar uma determinada solução, cada solução é representada como uma sequência de números (referido como um cromossomo). Após cada rodada de testes ou simulação, a ideia é excluir os "n" de piores soluções de design, e produzir "n" novos das melhores soluções. O AG gera um vetor com 30 posições em que os valores são 0 e 1. Assim as posições com o valor 1 significam as entradas selecionadas, são elas que serão utilizadas no treinamento do algoritmo de classificação, salvando sempre o melhor resultado.

A Tabela 6 mostra a lista de variáveis de entrada na primeira coluna e nas demais colunas mostra quais variáveis foram selecionadas por cada método de classificação utilizado (células assinaladas). Para o classificador KNN, nove variáveis das 30 apresentadas ao algoritmo foram selecionadas pelo AG, já para o classificador AD foram selecionadas 13 variáveis, e por fim para o algoritmo RNA, foram selecionadas 10 entradas.

Tabela 6 Variáveis selecionadas pelo AG

Variáveis de Entrada	Algoritmos		
	KNN	AD	RNA
Sexo			
Idade (anos)			
Estado civil		X	
Ocupação		X	
Educação		X	
Tempo de diagnóstico de diabetes (anos)			
Diabetes Tipo 1 ou 2			
Tempo de tratamento da diabetes (anos)			
Tipo de tratamento	X	X	
Hábito de examinar os pés			
Como corta as unhas dos pés	X	X	X
O que usa para lavar os pés		X	X
Hábito de lavar os pés	X	X	X
O que usa para limpar os pés	X		
Limpa entre os dedos			X
Usa creme hidratante nos pés		X	X
O que usa para remover calos			
Utiliza bolsa de água quente nos pés			X
Remove cutículas dos dedos dos pés			
Verifica o calçado dentro antes de usar	X	X	X
Tipo de sapato usado	X		
A mudança de atitude quando percebe alteração nos pés	X	X	
Aspecto interno do calçado		X	
Hora do dia para comprar sapatos novos		X	
Normalmente caminha com os pés descalços			
Tipo de meias que utiliza (material)			X
Tipo de meias que utiliza (aparência e costura)			

"Tabela 6, conclusão"

Variáveis de Entrada	Algoritmos		
	KNN	AD	RNA
Posição para assistir televisão	X		X
Material do sapato que usa			
Já percebeu mudanças nos pés	X	X	
Já foi diagnosticado com pé diabético			X

Fazendo-se a seleção de variáveis com AG, para os três classificadores considerados, foram selecionadas três entradas em comum, que são: como cortar as unhas dos pés, hábito de lavar os pés e verificar o calçado dentro antes de usar. Seis das 30 entradas aparecem na seleção de dois classificadores. Pode-se entender que, para o banco de dados considerado, estas nove variáveis são fundamentais para definir a qual grupo de risco o paciente pertence e, portanto, elas devem ser observadas com bastante atenção pela equipe de saúde local.

Tabela 7 Classificação utilizando AG

Classificador	Número de entradas selecionados	Nível de risco	Número de pacientes		Acurácia
			Cluster 1	*Cluster* 2	
KNN	9	Alto risco Baixo risco	15 0	0 15	1,00±0,00
AD	13	Alto risco Baixo risco	14 1	1 14	0,93±0.09
RNA	10	Alto risco Baixo risco	11 0	4 15	0,87±0,12

A Tabela 7 mostra a classificação dos 30 dados simulados utilizando o AG. Percebe-se uma melhora considerável, levando o classificador KNN de 7% para 100% de acerto, com nove entradas. O AD e o RNA também obtiveram

melhora considerando somente 13 e 10 entradas respectivamente, passando de uma acurácia de 0,57 ± 0,17 para 0,93 ± 0.09 (AD) e de 0,53 ± 0,17 para 0,87 ± 0,12 (RNA).

4.3.4 Classificação especialista

De acordo com o especialista, a classificação desses 40 pacientes é uma tarefa difícil, já que o número de variáveis é alta, o que faz com que a decisão final seja um problema multidimensional. A Tabela 8 mostra a classificação do especialista, na qual nota-se que 8 dos 40 pacientes não foram atribuídos a uma classe devido a não certeza do especialista, apoiando a afirmação de que esta é uma tarefa difícil.

Tabela 8 Classificação dos dados selecionados pela especialista

Grupos	A classificação do especialista			Acurácia
	Alto risco	Baixo risco	Não atribuído	
Alto risco	11	4	5	0,69 ± 0,16
Baixo risco	6	11	3	

Além disso, comparando a classificação do especialista com a classificação do método proposto, observou-se que 22 pacientes foram classificados de acordo com o método proposto e 10 foram classificados em desacordo, levando a uma acurácia de 0,69 ± 0,16.

Foi utilizado o *silhouttes* (Figura 17) para comparar o agrupamento da especialista com o agrupamento obtido pelo *k-means*.

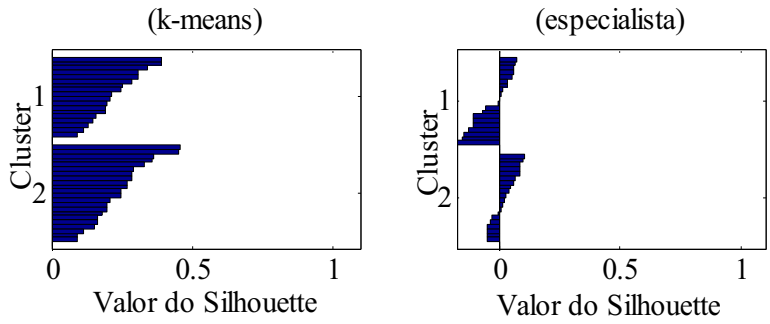


Figura 17 Agrupamento dos 40 dados pelo *k-means* e especialista

Percebe-se que o valor do *silhouettes* do agrupamento do *k-means* é maior que o do especialista, que obteve valores negativos, ou seja, os dados não foram agrupados adequadamente, de acordo com o critério *silhouettes*. Isso reafirma a dificuldade em agrupar os pacientes, pois os formulários contêm um número alto de variáveis.

5 CONCLUSÕES

A identificação oportuna de casos do pé diabético favorece o estabelecimento de vínculos entre os pacientes e as Unidades Básicas de Saúde, o que é imprescindível para o sucesso do controle do diabetes.

Entre as vantagens do acompanhamento e controle do diabetes, no âmbito da atenção básica, estão a possibilidade de evitar o surgimento de complicações, assim como de evitar o agravamento destas ocorrências. É consenso geral entre endocrinologistas e infectologistas que a principal medida no tratamento do pé diabético é a detecção precoce, alcançando-se mais de 90% de sucesso para as úlceras que recebem manejo adequado, incluindo alívio da pressão local, tratamento das infecções e revascularização quando indicada.

Este trabalho propôs um método automático não invasivo para identificar pacientes diabéticos que possuem alto risco de desenvolver o pé diabético. Usando dados reais e simulados, o método obteve resultados satisfatórios alcançando um desempenho de 100% para dados simulados e 68%, considerando a classificação dos especialistas como o padrão-ouro, para dados reais.

O método requer um processamento computacional simples e, portanto, pode ser implementado em computadores básicos em Unidades Básicas de Saúde, permitindo o controle automático e sistemático de pacientes diabéticos em relação à complicação do pé diabético. Além disso, o método pode oferecer um suporte aos agentes de saúde fazendo a triagem de pacientes com elevado risco de desenvolvimento do pé diabético.

O monitoramento dos pacientes por meio do método é importante para ver se um paciente classificado como alto risco vai migrar para o grupo de baixo risco após o acompanhamento sistemático da equipe de saúde.

6 PERSPECTIVAS FUTURAS

Como propostas futuras, que dependerão de novas parcerias, destacam-se:

a) Desenvolver um formulário eletrônico para coleta de informações;
b) Desenvolver uma plataforma amigável que implemente o método proposto em um microcomputador com uma configuração básica a ser utilizado em Unidades Básicas de Saúde;
c) Incorporar novas variáveis de entrada ao sistema, tais como informações sobre o nível de glicose do paciente, tabagismo e histórico de exames realizados ao longo do ano, de acordo com a opinião de um profissional clínico especializado;
d) Implementar o método em Unidades Básicas de Saúde para ser testado de forma mais efetiva. Nesta fase, um ponto importante será verificar se um paciente uma vez classificado com alto risco migrará para o *cluster* de baixo risco após o acompanhamento sistemático da equipe de saúde.

REFERÊNCIAS

AHAH, D.; KIBLER, D.; ALBERT, M. **Instance-based learning algorithms**. Boston: Machine Learning, 1991.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge: MIT Press, 2004.

ANTONELLI, D. et al. **Analysis of diabetic patients through their examination history**. Italia: Expert Systems with Applications, 2013. Não Publicado.

BELONGIE, S.; MALIK, J.; PUZICHA, J. Shape matching and object recognition using shape contexts. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 24, n. 4, p. 509–522, Apr. 2002.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. **Statistics for experimenters.** New York: John Wiley & Sons, 1978.

BOYKO, D. J. et al. Prediction of diabetic foot ulcer occurrence using commonly available clinical information: the Seatle diabetic foot study. **Diabetes Care**, Alexandria, v. 29, n. 6, p. 1202-1207, 2006.

BRASIL. Ministério da Saúde. Secretaria de Políticas de Saúde. Departamento de Ações Programáticas Estratégicas. **Plano de reorganização da atenção à hipertensão arterial e ao *diabetes mellitus*:** hipertensão arterial e *diabetes mellitus*. Brasília: Ministério da Saúde, 2002.

BREIMAN, L. et al. **Classification and regression trees**. Belmont: Wadsworth, 1984.

CASTANHEIRA, L. G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões.** 2008. 95 p. Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

CELEUX, G.; GOVAERT, G. Gaussian parsimonious *cluster*ing models. **Pattern Recognition**, Ezmsford, v. 28, n. 5, p. 781–793, May 1995.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, New York, v. 13, n. 1, p. 21–27, Jan. 1967.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via theEM algorithm. **Journal of the Royal Statistical Society**, Serie B, London, v. 39, n. 1, p. 1-38, 1977.

DISSAT, C.; RODRIGUES, P. Pé diabético. **Diabetes:** a revista da SBD, São Paulo, v. 20, n. 1, p. 11-14, jan. 2013.

DODGE, Y. **The concise encyclopedia of statistics**. New York: Springer, 2008.

DUARTE, N.; GONCALVES, A. Pé diabético. **Angiologia e Cirurgia Vascular**, Rio de Janeiro, v. 7, n. 2, p.65-79, jun. 2011.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2. edn. New York: John Wiley & Sons, 2000.

EKOE, J. et al. **The epidemiology of *diabetes mellitus***. 2. ed. Oxford:Wiley-Blackwell, 2008.

FARJADO, C. A importância do cuidado com o pé diabético: ações de prevenção e abordagem clínica. **Revista Brasileira Medicina da Família e Comunidade**, Rio de Janeiro, v. 2, n. 5, p. 43-58, abr./jun. 2006.

FERRAZ, A. E. P. et al. Atendimento multiprofissional ao paciente com *diabetes mellitus* no Ambulatório de Diabetes do HCFMRP-USP. **Medicina**, Ribeirão Preto, v. 33, p. 170-171, abr./jun. 2000.

FERREIRA, A. C. B. H. et al. Análise quantitativa do conhecimento do paciente diabético sobre o auto cuidado como prevenção do pé diabético. In: CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UFLA, 24., 2011, Lavras. **Anais...** Lavras: Editora da UFLA. (Resumo apresentado na forma de pôster).

FERREIRA, D. D.; SEIXAS, J. M. Qualidade de dados via árvores de decisão em apoio ao diagnóstico da tuberculose pulmonar. In: CONGRESSO BRASILEIRO DE AUTOMAÇÃO, 18., 2010, Mato Grosso do Sul. **Anais...** Mato Grosso do Sul: CBA, 2010.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. Desvendando os mistérios do coeficiente de correlação de Pearson (r)**. Revista Política Hoje**, Pernambuco, v. 18, n. 1, p. 115-146, 2009.

FORGY, E. *Cluster* analysis of multivariate data: efficiency versus interoperability of classification. **Biometrics**, Washington, v. 21, p. 768-769, 1965.

FRALEY, C.; RAFTERY, A. E. Model-based *cluster*ing, discriminant analysis, and density estimation**. Journal of the American Statistical Association**, New York, v. 97, n. 458, p. 611-612, June 2002.

FRANCO, L. J. et al. Diabetes como causa básica ou associada de morte no Estado de São Paulo**,** Brasil, 1992. **Revista de Saúde Pública**, São Paulo, v. 32, n. 3, p. 237-245, jun. 1998.

FREI, F. **Introdução à análise de agrupamentos**. São Paulo: Editora da UNESP, 2006.

GARSON, G. D. **Statnotes:** topics in multivariate analysis. [S.l.: s.n], 2006. Disponível em: <http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm>. Acesso em: 15 mar. 2013.

GIGANTE, D. P.; ASSUNÇÃO, M. C. F.; SANTOS, I. S. Atenção primária em diabetes no Sul do Brasil: estrutura, processo e resultado. **Revista de Saúde Pública**, São Paulo, v. 35, n. 1, p. 88-95, fev. 2001.

GOULARTE, A. et al. Redes neurais artificiais aplicadas no estudo de questionário de varredura para conjuntivite alérgica em escolares. **Arquivos Brasileiros de Oftalmologia**, São Paulo, v. 69, n. 5, p.707-713, set./out. 2006.

HAMERLY, G.; ELKAN, C. Alternatives to the *K-means* algorithm that find better *cluster*ings. In: PROCEEDINGS OF THE ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 2002, New York. **Anais…** New York: CIKM, 2002. p. 600-607.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. Burlington: M. Kaufmann, 2005.

HARRIS, M. I. Diabetes in America: epidemiology and scope of the problem. **Diabetes Care**, New York, v. 21, supl. 3, p. 11-14, 1998.

HAYKIN, S. **Neural networks and learning machines**. 3 ed. London: Prentice Hall, 2008.

HOLLAND, J. H. **Adaptation in natural and artificial systems.** Cambridge: University of Michigan Press, 1975.

IZENMAN, A. J. **Modern multivariate statistical techniques:** regression, classification, and manifold learning. New York: Springer, 2008.

KAREGOWDA, A.; JAYARAM, M.; MANJUNATH, A. Cascading *K-means cluster*ing and K-nearest neighbor classifier for categorization of diabetic patients. **International Journal of Engineering and Advanced Technology**, Oxford, v. 1, n. 3, p. 147-151, Feb. 2012.

KING, H.; AUBERT, R. E.; HERMAN, W. H. Global burden of diabetes, 1995-2025. **Diabetes Care**, New York, v. 21, n. 9, p. 1414-1431, Sept. 1998.

LEE, B. Y. et al. The economic effect of screening orthopedic surgery patients preoperatively for methicillin- resistant Staphylococcus aureus. **Infection Control and Hospital Epidemiology**, Chicago, v. 32, n. 11, p. 1130-1138, Nov. 2010.

LIN, S.-W.; CHEN, S.-C. **Parameter determination and feature selection for C4.5 algoritm using scatter search approach.** New York: Springer-Verlag, 2011.

LINDEN, R. **Algoritmos genéticos:** uma importante ferramenta da inteligência computacional. 2. ed. São Paulo: Brasport Livros e Multimídia, 2008.

LLOYD, P. Least square quantization in PCM. **IEEE Transactions on Information Theory**, New York, v. 28, n. 2, p. 129-137, Mar. 1982.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, California. **Proceedings…** California: University California Press, 1967. p. 281–297.

MALERBI, D. A.; FRANCO, L. J. Multicenter study of the prevalence of *diabetes mellitus* and impaired glucose tolerance in the urban Brazilian population aged 30-69 Yr. **Diabetes Care**, New York, v. 15, n. 1, p. 1509- 1516, Nov. 1992.

MATH WORKS. ***K-means Cluster*ing**. Torrance: The MathWorks, 2013. Disponível em: <http://www.mathworks.com/help/stats/*K-means-cluster*ing.html#bq_679x-19>. Acesso em: 24 mar. 2013.

MAYR, E. **Methods and principles of systematica zoology**. New York: McGraw-Hill, 1953.

MOORE, D. S. **The basic practice of statistics**. New York: Palgrave Macmilan, 2007.

OBITKO, Marek. **Introdução aos algoritmos genéticos.** Dresden: Universidade de Ciências Aplicadas, 2004. Disponível em: <http://www.obitko.com/ tutorials/genetic-algorithms/portuguese/dna-pictures.php>. Acesso em: 01 fev. 2013.

O NEURÔNIO artificial. Disponível em: <http://www.gsigma.ufsc.br/~ popov/aulas/rna/ neuronio_artificial/index.html>. Acesso em: 25 jul. 2013.

PANG-NING, T.; STEINBACH, M.; KUMAR, V. **Introdução à mineração de dados**. New York: Addison-Wesley, 2006.

PANDRE, A. **Análise de Cluster.** 2014. Disponível em: <http://apandre.wordpress.com/visible-data/cluster-analysis/>. Acesso em: 25 jul. 2013.

PATIL, B.; JOSHI, R.; TOSHNIWAL, D. Hybrid prediction model for Type-2 diabetic patients. **Expert Systems with Applications**, India, v. 37, n. 12, p. 8102–8108, Dec. 2010.

QUINLAN, J. **C4.5 programas de aprendizagem de máquina**. San Mateo: Morgan Kaufmann Publishers, 1993.

QUINTAL, G.M.C.C. **Análise de clusters aplicada ao Sucesso/Insucesso em Matemática**. 2006. Disponível em: <http://digituma.uma.pt/bitstream/10400.13/224/1/GuidaCaldeiraMestrado.pdf>. Acesso em: 25 jul. 2013.

REYNOL, F. **Mineração de dados para diagnósticos médicos**. São Paulo:FAPESP, 2010. Disponível em: < http://agencia.fapesp.br/11928>. Acesso em: 24 mar. 2013.

REYNOLDS, D. A. Robust text-independent speaker identification using gaussian mixture models. **IEEE Transactions on Speech and Audio Processing**, New York, v. 3, n. 1, p. 72-83, Jan. 1995.

REYNOLDS, D. **Gaussian mixture models**. Lexington: MIT Lincoln Laboratory, 2007.

RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: the rprop algorithm .In: IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, 1993, San Francisco. **Proceedings…** San Francisco: IEEE, 1993. p. 586–591.

ROSENBLANTT, F. **Principles of neurodynamics.** New York: Spartan Books, 1962.

ROUSSEEUW, P. *Silhouettes*: a graphical aid to the interpretation and validation of *cluster* analysis. **Journal of Computational and Applied Mathematics**, Antwerpen, v. 20, p. 53-65, Nov. 1987.

SANTANA, A. L. **Projeto e implementação de um sistema de suporte à decisão para o observatório de saúde da Amazônia**. 2005. 62 p. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Pará, Pará, 2005.

SANTOS, A. et al. Árvore de classificação e redes neurais artificiais: uma aplicação a predição e tuberculose pulmonar. In: Conference on Neural Networks, 6.; Congresso Brasileiro de Redes Neurais, 6., 2003, São Paulo. **Proceedings...** São Paulo, 2003. p. 427-432.

SANTOS, A. J.; ROSSI, V. E. C.; NASCIMENTO, E. Práticas utilizadas no uso de insulina em domicílio. **Ciência et Praxis**, Passos, v. 3, n. 5, p. 43-46, jan./jun. 2010.

SANTOS, A. J.; ROSSI, V. E. C.; OLIVEIRA, M. L. Conhecimento do paciente diabético em relação à autoaplicação de insulina e descarte apropriado de materiais perfurocortantes. **Nursing**, São Paulo, v. 13, n. 155, p. 209-213, abr. 2011.

SANTOS, A. M. et al. Usando redes neurais artificiais e regressão logística na predição da hepatite A. **Revista Brasileira de Epidemiologia**, São Paulo, v. 8, n. 2, p. 117-126, 2005.

SARTORELLI, D. S.; FRANCO, L. J. Tendências do *diabetes mellitus* no Brasil: o papel da transição nutricional. **Caderno de Saúde Pública**, Rio de Janeiro, v. 19, supl. 1, p. 529-536, 2009.

SASLAW, W. C. Some properties of a statistical distribution function for galaxy *cluster*ing**. Astrophysical Journal**, Chicago, v. 341, p. 588-598, June 1989.

SCAPIM, E. P. **Perfil dos pacientes com *diabetes mellitus* que possuem úlcera no pé, atendidos em unidade ambulatorial da cidade de Marília-SP**. 2004. 157 p. Ribeirão Preto, Dissertação (Mestrado em Enfermagem) - Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, 2004.

SCHEFFEL, R. S. et al. Prevalência de complicações micro e macrovasculares e de seus fatores de risco em pacientes com diabetes melitodo tipo 2 em atendimento ambulatorial. **Revista da Associação Medica Brasileira, São Paulo**, v. 50, n. 3, p. 263-267, jul./set. 2004.

SILBERSCHATZ, A. et al. **Sistema de banco de dados**. 5. ed. Rio de Janeiro: Elsevier, 2006.

SIMARD, P.; LECUN, Y.; DENKER, J. Efficient pattern recognition using a new transformation distance. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 5., 1992, San Francisco. **Proceeding…** San Francisco: Morgan Kaufmann Publishers, 1992. p. 50–58.

SNEATH, P. H.; SOKAL, R. R. **Numerical taxonomy:** the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973.

SOUZA, B. et al. **Redes Neurais aplicadas ao diagnóstico da tuberculose pulmonar paucibacilar**. Rio de Janeiro: Editora da UFRJ, 2007.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining:** mineração de dados. São Paulo: Ciência Moderna, 2009.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Addison-Wesley, 2005.

TORRES, H. C. et al. Avaliação estratégica de educação em grupo e individual no programa educativo em diabetes. **Revista de Saúde Pública**, Belo Horizonte, v. 43, n. 2, p. 291-298, 2009.

TURI, R. H. ***Cluster*ing-based colour image segmentation**. 2001. 373 p. Thesis (PhD) - Monash University, Australia, 2001.

VIGO, K. O.; NUNES, P. D.; PACE, A. E. O conhecimento dos familiares acerca da problemática do portador de *diabetes mellitus*. **Revista Latino-Americana de Enfermagem**, Ribeirão Preto, v. 11, n. 3, p. 312-319, maio/jun. 2003.

ZHUANG, X. et al. Gaussian mixture density modeling: decomposition, and applications. **IEEE Transactions on Image Processing**, New York, v. 5, n. 9, p. 1293-1302, Sept. 1996.

ZUBEN, F. V. **Computação evolutiva:** uma abordagem pragmática. São Paulo: Editora da UNICAMPI, 2004. Disponível em: <ftp://ftp.dca.fee.unicamp.br/ pub/docs/vonzuben/ ia707_1s04/ textos/tutorialEC.pdf>. Acesso em: 01 fev. 2013.

APÊNDICE

APÊNDICE A - Lista de publicações

Nesta seção, as publicações realizadas em congressos serão apresentadas. Estes estão organizados em ordem cronológica.

Trabalhos completos publicados em anais de congressos

SILVA, R. N. ; FERREIRA, D. D. ; CARVALHO, V. A. ; BORGES, A. P. R. ; BARBOSA, B. H. G. ; FERREIRA, A. C. B. H. . **Identificação de Pacientes Diabéticos com Potencial para Desenvolver o Pé Diabético**. In: XI Congresso Brasileiro de Inteligência Computacional, 2013, Recife, Porto de Galinhas. Anais do XI Congresso Brasileiro de Inteligência Computacional, 2013. p. 1-6.

Resumos expandidos publicados em anais de congressos

SILVA, R. N. ; FERREIRA, D. D. ; BARBOSA, B. H. G. ; FERREIRA, A. C. B. H. . **Identificação de pacientes diabéticos com potencial para desenvolver o pé diabético: uma abordagem multidisciplinar**. In: xxii congresso de pós-graduação da ufla, 2013, lavras. Xxii congresso de pós-graduação da ufla, 2013.

Resumos publicados em anais de congressos

SILVA, R. N. ; FERREIRA, A. C. B. H. ; BARBOSA, B. H. G. ; FERREIRA, D. D. ; ROSSI, V. E. C. . **Seleção de Ações de Autocuidado relevantes à identificação de pacientes diabéticos com potencial para desenvolver o pé diabético.** In: Congresso brasileiro de enfermagem, 2013, Rio de Janeiro. 65º Congresso Brasileiro de Enfermagem, 2013.

ANEXOS

ANEXO A - Formulário

1-Sexo: ( ) F ( ) M
2-Idade:_____ anos
3-Estado Civil: ( ) Solteiro () Viúvo () Divorciado () Casado
4-Ocupação: _____
5-Escolaridade: ____ anos de estudo
6-Há quanto tempo tem diabetes? _____
7-Tipo de diabetes: ( )Tipo 1 ( ) Tipo 2 Outros_____
8-Tempo de tratamento do diabetes: _____
9-Tipo de tratamento: ( ) Antidiabético oral ( ) Insulina ( ) Dieta (Insulina: se sim, Quantas unidades?_____)

10-Você tem o hábito de examinar seus pés?

( ) sim ( ) não

Com que frequência?

( ) todos os dias ( ) uma vez por semana

( ) uma vez por mês ( ) uma vez por ano

( ) quando há necessidade

11-Como você corta suas unhas?

( ) rente ao dedo quadrada (reta)

( ) rente ao dedo redonda (cortando os cantos)

( ) não rente ao dedo redonda (cortando os cantos)

( ) não rente ao dedo quadrada (reta)

Se você não corta, quem faz?_____

12-O que você usa para lavar seus pés?

( ) sabão de coco ( ) sabonete comum

( ) sabonete neutro ( ) só água

( ) outro?_____

13-Você tem o hábito de lavar os pés com água e sabão, esfregando com bucha ou outro material todos os dias?

( ) sim ( ) não ( ) às vezes.

Qual frequência?_____

14-O que usa para enxugar os pés?

( ) toalha comum ( ) toalha macia

( ) toalha crespa ( ) pano de chão

( ) papel ( ) outro

15-Costuma enxugar entre os dedos todas as vezes que os pés ficam molhados?

( ) sim ( ) não ( ) às vezes

16-Costuma passar creme hidratante nos pés?

( ) sim. ( ) não

Local:

( ) entre os dedos e na sola do pé

( ) em cima e na sola do pé

( ) em cima, na sola e no calcanhar

( ) em cima, na sola, entre os dedos e no calcanhar

17-O que você costuma usar para remover calos?

( ) lixa de papel e creme hidratante

( ) lixa de metal e creme hidratante

( ) pedra-ume ou pedra-pomes e creme hidratante

( ) pedra normal e creme hidratante

( ) substância química (calicida)

( ) nunca teve calos

18-Quando sente seus pés frios, costuma usar bolsa de água quente?

( ) sim ( ) não ( ) às vezes

19-Tem o hábito de retirar cutículas?

( ) sim ( ) não ( ) às vezes

20-Que tipo de calçado você costuma usar?

( ) aberto ( ) fechado com costura

( ) fechado sem costura

21-Tem o hábito de verificar o calçado por dentro antes de usá-lo?

( ) sim ( ) não ( ) às vezes

22-Em relação ao material do calçado, qual você usa?

( ) de pano ( ) de couro ( ) de couro sintético

23-Como é o aspecto interno do seu calçado?

( ) sem costura ( ) com costura

( ) com e sem costura

24-A que horas você costuma sair para comprar sapatos novos?

( ) pela manhã ( ) qualquer hora

( ) início da tarde ( ) final da tarde

25-Você costuma andar descalço?

( ) só em casa ( ) em casa e na rua

( ) na rua ( ) nunca fica descalço

26-Qual tipo de meia você prefere usar?

( ) de algodão ( ) de fio sintético

( ) de lã ( ) outro_____

27-Que tipo de meias você usa?

( ) claras e com costura

( ) claras e sem costuras

( ) escuras e com costuras

( ) escuras e sem costuras

( ) escuras e claras sem costura

( ) escuras e claras com costura

( ) outro_____

28-Como você costuma assistir televisão?

( ) com os pés para baixo

( ) com as pernas cruzadas

( ) com as pernas elevadas

( ) fica deitado

29-Se você percebe alguma alteração nos pés, qual é sua atitude?

( ) procura médico/enfermeiro

( ) usa medicação (pomada, creme)

( ) coloca os pés para cima

( ) não faz nada, espera a evolução

30- Você já percebeu ou sentiu alguma alteração nos pés?

( ) sim ( )não

Qual? _____

31- Você já foi diagnosticado com o pé diabético?

( ) sim. ( ) não

Quando foi a última vez? _____

Quantas vezes? _____

ANEXO B - Termo de Consentimento Livre e Esclarecido TCLE

Nome:_____

As informações contidas neste termo visam firmar acordo por escrito, mediante o qual o responsável pelo menor ou o próprio sujeito objeto de pesquisa, autoriza sua participação, com pleno conhecimento da natureza dos procedimentos e riscos a que se submeterá, com capacidade de livre arbítrio e sem qualquer coação. O TCLE deve ser redigido em linguagem acessível ao voluntário de pesquisa.

I - TÍTULO DO TRABALHO EXPERIMENTAL:

Desenvolvimento de um Sistema Neural para a Identificação Precoce de Pacientes Diabéticos com Potencial para Desenvolver o Pé Diabético

Pesquisador Responsável: Prof. Dr. Danton Diego Ferreira

II - OBJETIVOS

Este projeto visa desenvolver um sistema automático utilizando redes neurais artificiais para a identificação precoce de pacientes que possuem potencial para desenvolver o pé diabético.

III - JUSTIFICATIVA

O sistema permitirá um controle mais dinâmico do problema do pé diabético e direcionará o paciente para campanhas mais específicas de acordo com a sua caracterização prévia dada pelo sistema, prevenindo assim o desenvolvimento do pé diabético.

IV - PROCEDIMENTOS DO EXPERIMENTO

AMOSTRA

Será feito uma visita à casa dos portadores de diabetes, em que algumas informações serão coletadas por meio de questionário aplicado por especialista. A identidade de cada paciente será preservada, de forma que suas informações serão utilizadas apenas para o projeto e teste do sistema proposto.

EXAMES

Não haverá exames, uma vez que a amostra deste trabalho será composta por pacientes diabéticos já cadastrados nas Unidades de Saúde Básica PSF.

V - RISCOS ESPERADOS

Dentro do conjunto amostral (número de pacientes diabéticos entrevistados) pode acontecer de haver um número inferior a 20% de pacientes que já desenvolveram e/ou apresentam o pé diabético. Esse pequeno número de casos positivos de pé diabético pode comprometer a eficiência do modelo neural. Assim, a solução imediata seria a replicação de alguns indivíduos para se alcançar os 20%. A solução não imediata seria aumentar o tamanho da amostra do trabalho.

VI – BENEFÍCIOS

Como benefícios destacam-se:

1) O desenvolvimento de uma ferramenta automática de baixo custo para o controle do pé diabético;

2) Permitirá a análise a fundo dos fatores mais causadores do pé diabético;

3) Fará a triagem de pacientes para o direcionamento a campanhas de autocuidado com os pés, facilitando o trabalho da equipe dos PSFs (Programa Saúde Família);

4) Atuará diretamente, com o apoio das equipes de saúde dos PSFs, na redução do número de casos de pé diabético;

5) Será um fator determinante no autocuidado do paciente quanto aos pés.

VII - RETIRADA DO CONSENTIMENTO

O responsável pelo menor ou o próprio sujeito tem a liberdade de retirar seu consentimento a qualquer momento e deixar de participar do estudo, sem qualquer prejuízo ao atendimento a que está sendo ou será submetido.

VIII – CRITÉRIOS PARA SUSPENDER OU ENCERRAR A PESQUISA

A princípio, critérios de ordem técnica para suspender ou encerrar a pesquisa são desconhecidos.

IX - CONSENTIMENTO PÓS-INFORMAÇÃO

PACIENTE MENOR DE IDADE

Eu_____,
responsável pelo menor
_____, certifico que, tendo lido as informações acima e suficientemente esclarecido (a) de todos os itens, estou plenamente de acordo com a realização do experimento. Assim, eu autorizo a execução do trabalho de pesquisa exposto acima.

Lavras, _____ de _____ de 20__.

NOME

(legível)_____RG_____

ASSINATURA_____

PACIENTE MAIOR DE IDADE

Eu_____

___, certifico que, tendo lido as informações acima e suficientemente esclarecido (a) de todos os itens, estou plenamente de acordo com a realização do experimento. Assim, eu autorizo a execução do trabalho de pesquisa exposto acima.

Lavras, _____ de _____ de 20__.

NOME

(legível)_____RG_____

ASSINATURA_____

ATENÇÃO: A sua participação em qualquer tipo de pesquisa é voluntária. Em caso de dúvida quanto aos seus direitos, escreva para o Comitê de Ética em Pesquisa em seres humanos da UFLA. Endereço – Campus Universitário da UFLA, Pró-reitoria de pesquisa, COEP, caixa postal 3037. Telefone: 3829-1127, falar com Andréa.

No caso de qualquer emergência, entrar em contato com o pesquisador responsável no Departamento de Engenharia. Telefones de contato: (35) 3829-1025.