USO DO ALGORITMO SIMPLECART COMO MÉTODO DE AUXÍLIO NA TOMADA DE DECISÃO EM SAÚDE

Karlenne Raquel de Brito Nascimento[1]
Ingrid Bergmam do Nascimento Silva[2]
Danilo Rangel Arruda Leite [3]
Mirian Marques Vieira[4]
Jozemar Pereira do Santos [5]

RESUMO

Buscando classificar características do paciente segundo a diabetes, foi utilizado o algoritmo SimpleCart para gerar árvores de decisão. Objetivou-se analisar a adequação do método de decisão SimpleCart no banco de dados do Instituto Nacional de Diabetes e Doenças Digestivas do Rim da Universidade Johns Hopkins, que trata da temática do diabetes em 1990. A base de dados utilizada foi constituída de 768 observações. O processo de mineração dos dados foi realizado através do software WEKA versão 3.8.0, a estratégia de validação do método consistiu em testar a forma de treino da amostra, utilizando o algoritmo CART, os dados foram divididos aleatoriamente em um conjunto de treinamento e um conjunto de teste, (50% de treinamento e 50% de teste (cross-validation). Foram variadas as porcentagens do tamanho da amostra do conjunto de treinamento (C) em 0,59, 0,85, 0,95, 0,97 e 0,98. Os k-flods do cross-validation variaram 2, 10, 20, 30, 40, 50 e 60. O Número mínimo de observações nos nós terminais (M) 1, 2, 3, 4, 5. E a sementes do número aleatório (S) testadas foram 1, 100, 200, 300, 400, 500 e 700. O índice Kappa, revelou níveis de concordância entre razoável e bom. O modelo inicial gerado pelo default do aplicativo Weka apresentou Kappa de 0,45 (M 2.0 / C 1.0/ S 1), considerado moderado. Após os testes o melhor Kappa obtido foi de 0,47 (M 4.0 / C 0.98 / S 200), indicando boa concordância. Concluiu-se que o algoritmo SimpleCart se mostrou computacionalmente eficiente, podendo auxiliar os gestores na tomada de decisão em saúde.

Palavras-chave: Diabetes, Weka, SimpleCart.

INTRODUÇÃO

O Diabetes *Mellitus* (DM) é uma doença caracterizada pela elevação da glicose no sangue (hiperglicemia). Pode ocorrer devido a defeitos na secreção ou na ação do hormônio insulina, que é produzido no pâncreas, pelas chamadas células beta. A diabetes está associada com danos a longo prazo em diferentes órgãos, especialmente olhos, rins, nervos, coração e vasos sanguíneos (PORTER; KAPLAN, 2014).

[1]Bióloga. Mestranda em Modelos de Decisão e Saúde na Universidade Federal da Paraíba - UFPB, karlenne_raquel@hotmail.com;
[2]Enfermeira. Mestranda em Modelos de Decisão e Saúde na Universidade Federal da Paraíba - UFPB, ingridgba2006@hotmail.com;
[3]Mestre em Informática. Doutorando em Modelos de Decisão e Saúde na Universidade Federal da Paraíba - UFPB, danilorangel@buscapb.com.br;
[4]Mestra em Modelos de Decisão e Saúde na Universidade Federal da Paraíba- UFPB, mirian_mvieira@hotmail.com;
[5]Doutor em Ciências da Saúde. Professor de graduação e pós-graduação da Universidade Federal da Paraíba - UFPB, jozemar@de.ufpb.br.

Estimativas mundiais confirmam que 382 milhões de pessoas vivem com a DM (8,3%), e que esse número poderá crescer e chegar a 592 milhões em 2035 (MOURA et al., 2012). Outro ponto importante, é que aproximadamente 50% dos diabéticos desconhecem que têm a doença. Até 2030, o DM pode ser considerada a sétima doença que mais causará mortes em todo mundo. Destaca-se, atualmente, como uma importante causa de morbidade e mortalidade (SHAW et al., 2010).

Os principais fatores considerados responsáveis pela prevalência e aumento da incidência do DM em todo o mundo são: o sedentarismo, o crescente aumento da obesidade, além do envelhecimento da população, e os processos de urbanização (GUARIGUATA et al., 2014).

Dessa maneira, levando em consideração a gravidade dessa doença, torna-se importante criar mecanismos capazes de classificar de maneira mais precisa indivíduos com determinadas características de risco para o desenvolvimento do diabetes.

O processo de descoberta de conhecimento em banco de dados é possível através de vários métodos, destacando-se os métodos de classificação e regressão que permitem obter padrões através da construção de árvores de decisão. Essas árvores são construções simples de entender, onde os modelos seguem uma estrutura gerada por regras compreensíveis interligadas por interações sucessivas e dependentes. Na forma de algoritmo de árvore de regressão, o tratamento de valores faltantes é satisfatório, comparado a modelo estatísticos clássicos (KALMEGH, 2015).

Buscando classificar características do paciente segundo essa doença, o fator diagnóstico do diabetes foi tratado como uma variável categórica binária, apresentando os nível positivo e negativo. Nesse sentido, foi utilizado o algoritmo SimpleCart para gerar árvores de decisão. O método SimpleCart representa a implementação do algoritmo *Classification And Regression Trees* (CART) proposto por Breiman em 1984 (BREIMAN et al, 1984), e consiste de uma técnica não-paramétrica que é usada para a exploração de dados, previsão e construção de árvores de classificação ou regressão, dependendo se o atributo é nominal (classificação) ou contínuo (regressão) (KHARCHE et al., 2014).

O algoritmo faz uso da Entropia para a escolha da melhor divisão entre todos os atributos. Além disso, essa metodologia usa validação cruzada ou testes de independência dos dados para selecionar a melhor árvore e a sequência dos galhos considerados no processo de decisão (SRINIVASAN; MEKALA, 2014).

Diante disso, objetiva-se analisar a adequação do método de decisão *SimpleCart* no banco de dados do Instituto Nacional de Diabetes e Doenças Digestivas do Rim da Universidade Johns Hopkins, que trata da temática do diabetes em 1990.

METODOLOGIA

A base de dados utilizada é constituída de 768 observações, onde a população é oriunda de Phoenix, no Arizona. Essa população é estudada desde 1965 pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais da Johns Hopkins University, por causa de sua alta incidência de diabetes. Além disso, todos os pacientes são mulheres com pelo menos de 21 anos de idade, de descendentes indígenas da tribo Pima (SMITH et al., 1988). Os dados foram obtidos por meio do seguinte endereço eletrônico http://repository.seasr.org/Datasets/UCI/arff/.

Como pode ser observado na Tabela 1 as oito variáveis independentes são todas do tipo intervalar e a variável de decisão é binária (teste positivo/teste negativo). Na divisão binária das variáveis contínuas é adotada uma técnica de divisão baseada na pesquisa exaustiva do ponto de divisão, essa realizada automaticamente pelo software.

O processo de mineração dos dados foi realizado através do software WEKA versão 3.8.0, pela utilização do algoritmo já implementado de arvore de classificação e regressão (CART), o SimpleCart.

A priori foi realizada uma análise descritiva do banco de dado e posteriormente a aplicação do SimpleCart.

Tabela 1: Variáveis independentes e variáveis de decisão da base de dados.

Codificação	Variáveis*	Escala de Mensuração
	Variáveis Independentes	
1-preg	Número de vezes que a mulher engravidou	Intervalar
2-plas	Concentração plasmática de glicose a 2 horas em um teste oral de tolerância à glicose.	Intervalar
3-pres	Pressão sanguínea diastólica (mm Hg).	Intervalar
4-skin	Espessura da dobra da pele do tríceps (mm).	Intervalar
5-insu	Nível de insulina sérica de 2 horas (mu U / ml).	Intervalar
6-mass	Índice de massa corporal (kg /m^2).	Intervalar

7-pedi	Função de qualidade de vida do diabetico.	Intervalar
	Variável de decisão	
8-age	Idade das mulheres em anos.	Intervalar
Class	teste positivo/ teste negativo	Categórica

*Descrição original da base

SimpleCart

O SimpleCart foi aplicado com a finalidade de analisar a qualidade e fidedignidade quanto a tomada de decisão. O critério utilizado para a divisão baseia-se na construção de uma árvore de decisão, sendo que seu primeiro passo é descobrir qual dos atributos realiza a melhor divisão. O algoritmo CART considera três critérios possíveis para selecionar a melhor divisão dos dados: Entropia, Critério de Twoing e Critério de Gini (AHER; LOBO, 2012).

As árvores geradas pelo algoritmo CART são sempre binárias as quais podem ser percorridas da sua raiz até as folhas respondendo apenas a questões simples do tipo “sim” ou “não”. Além disso, dispõe de um tratamento especial para atributos ordenados e também permite a utilização de combinações lineares entre atributos (agrupamento de valores em vários conjuntos), sendo de fácil interpretação dos resultados, já que a classificação é obtida de forma explícita (TRUJILLANO et al., 2008).

Essa abordagem é uma maneira completamente diferente de particionamento. São compostas por nós (raiz, filho e terminal) e ramos, iniciando sempre no nó principal, seguido dos nós filhos (intermediários), sendo ligados por ramos e finalizada pelo nó terminal ou folha (Figura 1). Além do fato que dispõe de um tratamento especial para atributos ordenados e também permite a utilização de combinações lineares entre atributos (agrupamento de valores em vários conjuntos), produzindo árvores mais simples, precisas e com boa capacidade de generalização (AITKENHEAD, 2008).

Para essa construção é necessário: definir o conjunto de regras para dividir cada nó da árvore; decidir quando a árvore está completa; associar cada nó terminal a uma classe ou a um valor preditivo. No caso da regressão, o algoritmo é usado para a exploração de dados, previsão e classificação das árvores de regressão, fazendo uso de dados históricos. Além disso, o CART gera amostras com valores de relevância para cada “galho” da árvore de decisão binária (DHAKATE et al., 2014).

A separação dos dados ocorre até que cada subconjunto esteja homogêneo, com casos de uma única classe (AHER; LOBO, 2012). Os nós que estão localizados abaixo do nó raiz são os nós filhos e esses nós estão conectados por ramos. As folhas são as regiões que estão associadas a um rótulo ou valor. Para dividir um conjunto em dois subgrupos, o algoritmo sempre faz perguntas dicotômicas, ou seja, que tem apenas um "sim" ou um "não" como resposta. Por exemplo, as questões podem ser: a idade é <=45? ou o crédito é <=500?

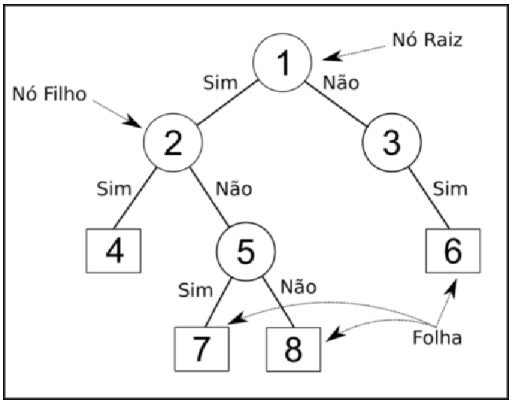

Figura 1: Esquema de árvore de decisão CART

O próximo passo do SimpleCart é ordenar cada regra de divisão com base no critério de qualidade. O critério padrão usado para classificação é o Índice de Gini que tem por base o cálculo da entropia (DHAKATE et al., 2014).

$$\phi(p_1,\ldots,p_2) = -\sum_j p_j \log p_j$$

em que p é a frequência encontrada de cada classe j, e o processo de divisão da árvore de regressão procura minimizar R(T).

$$R(T) = \frac{1}{N} \sum_{t \epsilon T} \sum_{x \epsilon t} (y - \bar{y}(t))^2$$

sendo t o identificador de cada nó da árvore e R(T) o valor esperado da soma dos erros quadráticos da regressão utilizando uma constante como modelo preditivo (a média).

O CART não apresenta na árvore de regressão, um modelo linear em seus nós terminais e sim uma média Y_s em cada partição. Uma vez encontrada a melhor divisão, repete-se o processo de procura para cada nó filho, continuamente até que a divisão seja impossível ou interrompida (SMITH et al., 1988).

A metodologia de regressão CART é desenvolvida em três etapas: I- o crescimento da árvore procedendo a diversas ramificações binárias no sentido de diminuir a diversidade da variável em estudo; II- validação da árvore; III- a interpretação da árvore de regressão proposta, na qual o papel da medida de importância relativa das variáveis deverá ser tido em conta (KALMEGH, 2015).

A estratégia de validação do método consistiu em testar a forma de treino da amostra. Nesse sentido, foi fixada a estratégia de teste cross-validation, variando os percentuais do banco de treinamento, pois é a mais utilizada na literatura.

Foi utilizado o parâmetro do coeficiente kappa que é uma medida de concordância entre juízes bastante utilizada para avaliar a qualidade de classificadores em classificar de maneira corretas os atributos nas respectivas classes. O resultado obtido pelo coeficiente Kappa, varia no intervalo de 0 a 1, sendo que quanto mais próximo a 1, melhor a qualidade dos dados classificados. Esses valores podem ser interpretados da seguinte maneira: concordância pobre com valores menores que 0,20; concordância pequena com valores entre 0,20 e 0,40; concordância moderada com valores entre 0,41 e 0,60; concordância boa com valores entre 0,61 e 0,80; concordância muito boa para valores oscilando entre 0,80 e 1 (KOTZ; JOHNSON, 1983).

Também foi analisada a matriz de confusão, que proporciona uma medida efetiva do modelo de classificação, apontando o número de classificações corretas versus as classificações preditas para cada classe, sobre um conjunto de variáveis. O número de acertos para cada classe se localiza na diagonal principal da matriz e os demais componentes representam os erros de classificação (AHER; LOBO, 2012).

DESENVOLVIMENTO

A diabetes mellitus (DM) é uma doença crônica que se desenvolve quando o pâncreas não pode produzir insulina suficiente ou quando há incapacidade da insulina de exercer adequadamente seus efeitos. As complicações do DM podem acontecer das seguintes formas: agudas com a hiperglicemia e a hipoglicemia e as crônicas que podem ser macrovasculares (doença cardíaca coronária, doença vascular periférica e doença cerebrovascular),

microvasculares (retinopatia e nefropatia) e neurológicas (neuropatia) (BEAGLEY et al, 2014).

O Diabetes se enquadram em duas categorias: Diabetes tipo 1, a causa é uma deficiência absoluta de secreção de insulina. Na outra, Diabetes tipo 2, a causa é uma combinação de resistência à ação da insulina e uma resposta secretiva inadequada à secreção de insulina. Na última categoria, um grau de hiperglicemia suficiente para causar alterações patológicas e funcionais em vários tecidos alvo, mas sem sintomas clínicos, pode estar presente por um longo período de tempo antes do diabetes ser detectado. Durante este período assintomático, é possível demonstrar uma anormalidade no metabolismo de carboidratos, medindo a glicemia no estado de jejum ou após um desafio com uma carga oral de glicose (GUARIGUATA et al., 2014).

Os modelos derivados de árvores de decisão, são facilmente constituídos por um fluxograma com estrutura de árvore e facilmente convertidas em regras de classificação (SPEYBROECK, 2012).

A abordagem CART é uma alternativa aos métodos tradicionais para predição. Dentro de sua implementação, o conjunto de dados é dividido em dois subgrupos que são os mais diferentes possíveis com respeito aos resultados. Este procedimento é continuado em cada subgrupo até que algum subgrupo de tamanho mínimo é atingido. As principais características deste algoritmo é a capacidade de generalização, trabalhar com variáveis quantitativas e qualitativas e possuir apenas decisões binária em todos os nós (SRINIVASAN; MEKALA, 2014).

O software WEKA (Waikato Environment for Knowledge Analysis) é um software de código aberto emitido sob a GNU General Public License, desenvolvido pela Universidade de Waikato na Nova Zelândia. Um dos maiores diferenciais do WEKA, é ser um software robusto em relação a recursos, é completo em termos de algoritmos e aplicações, bem como a sua capacidade de ser incorporado como built-in em outras aplicações através da biblioteca Java. Fornece uma interface uniforme para muitos algoritmos de aprendizagem diferentes, juntamente com métodos para pré e pós processamento e para avaliar o resultado dos esquemas de aprendizagem em qualquer conjunto de dados. Ele também inclui uma variedade de ferramentas para transformar conjuntos de dados, como *os algoritmos, o*s principais algoritmos que estão disponíveis são: regressão, classificação, mineração de regras de associação, seleção de atributos, e entre outros (DHAKATE, 2014).

A implementação do algoritmo CART é denotada por SimpleCart, no WEKA, sendo baseado em métodos proposto por Breiman, Friedman, Oslen e Stone em 1984. Dentre as principais características do algoritmo está a grande capacidade de pesquisa de relações entre os dados, mesmo quando elas não são evidentes, bem como a produção de resultados sob a forma de árvores de decisão de grande simplicidade e legibilidade (SPEYBROECK, 2012).

RESULTADOS E DISCUSSÃO

Como primeiro passo na análise do método foi necessário fazer a análise descritiva dos dados da diabetes. Na Tabela 2 observa-se as estatísticas descritiva dos dados, sendo possível perceber que o atributo de concentração plasmática de glicose a 2 horas em um teste oral de tolerância a glicose, apresentou uma média de 120,90, com desvio padrão 31,97. O atributo 1, referente ao número de vezes que a mulher engravidou, possui média de 3,845 com desvio padrão de 3,37.

Essa análise dos dados mostra que existem problemas em relação a base de dados, com observações faltantes, pois a maioria das variáveis não permite que seja observado o valor 0, como por exemplo a pressão sanguínea. Os altos valores de desvio-padrão indicam a grande variabilidade nas informações coletadas.

Tabela 2: Estatísticas descritiva.

Atributo	Mínimo	Máxima	Média	Desvio Padrão
1	0,0	17,0	3,845	3,37
2	0,0	199,0	120,90	31,97
3	0,0	122,0	69,11	19,36
4	0,0	99,0	20,54	15,95
5	0,0	846,0	79,80	115,24
6	0,0	67,1	31,99	7,88
7	0,1	2,4	0,47	0,33
8	21,0	81,0	33,24	11,76

A literatura aborda que o exame padrão utilizado para o rastreamento do DM é o teste de tolerância à glicose oral (TTGO) de duas horas (SALLEY et al, 2007) Entretanto, existem Controvérsias em relação a esse teste, como a duvida se este deve ser realizado em todos os pacientes, somente em obesos, ou na presença de fatores de risco como a idade ≥45 anos,

obesidade, adiposidade abdominal, resistência à insulina e história familiar de DM (MORAN et al, 2011).

Utilizando o algoritmo CART, os dados foram divididos aleatoriamente em um conjunto de treinamento e um conjunto de teste, (50% de treinamento e 50% de teste (cross-validation). Para atingir os objetivos foram variadas as porcentagens do tamanho da amostra do conjunto de treinamento (C) em 0,59, 0,85, 0,95, 0,97 e 0,98. Os k-flods do cross-validation variaram 2, 10, 20, 30, 40, 50 e 60. O Número mínimo de observações nos nós terminais (M) 1, 2, 3, 4, 5. E a sementes do número aleatório (S) testadas foram 1, 100, 200, 300, 400, 500 e 700.

Esses parâmetros foram testados em conjunto, sendo avaliados os índices Kappas de cada modelo. Como resultado dessa avaliação o índice Kappa, revelou níveis de concordância entre razoável e bom. O modelo inicial gerado pelo default do aplicativo Weka apresentou Kappa de 0,45 (M 2.0 / C 1.0/ S 1), considerado razoável. Após os testes o melhor Kappa obtido foi de 0,47 (M 4.0 / C 0.98 / S 200), indicando boa concordância.

Tabela 3- Medidas resumo do ajuste do CART

Medidas*	
Classificação Correta	77%
Classificação Incorreta	23%
Kappa	0,47
Erro médio	0,31
Precisão	76%
Curva ROC	76%

*M 4.0 / C 0.98 / S 200

O trabalho de Bhavsar (2016) faz a comparação de diferentes métodos de classificação para o mesmo banco de dados analisado (diabetes), encontrando os valores de Kappa nos mesmo patamar de concordância que para o SimpleCart, nesse sentido há evidencias que a precisão da previsão dos modelos testados nesse é afetada pelo conjunto de dados.

No algoritmo testado as divisões dos nós são binárias, nesse sentido, observe no esquema abaixo a árvore do modelo final gerada pelo Software Weka, com 5 nós. Note que as árvores foram construídas com apenas duas variáveis. As outras variáveis podem ter sido excluídas da arvore por apresentarem pequena correlação com a varável desfecho (class).

Tabela 4- Árvore de decisão gerada pelo SimpleCart

CART Decision Tree

plas < 127.5: tested_negative(391.0/94.0)

plas >= 127.5

| mass < 29.95: tested_negative(52.0/24.0)

| mass >= 29.95: tested_positive(150.0/57.0)

*Output Software Weka 3.8.1

De modo geral, a principal característica do algoritmo é a grande capacidade de pesquisa de relações entre os dados, independente da evidencia dessas relações, bem como, tem forte influência na produção e geração de resultados sob a forma de árvores de decisão, quase sempre de ordem binária, de grande simplicidade e legibilidade (TRUJILLANO et al, 2008).

Em relação a matriz de confusão observou-se que foram classificados corretamente como teste negativo 438 de 500 casos e como teste positivo 154 de 258. Em geral, nas tentativas de encontrar o melhor modelo, observou-se na matriz de confusão um comportamento semelhante, uma vez que, o acerto na classificação de teste negativo é maior que no teste positivo.

Tabela 5- Matriz de confusão gerada pelo SimpleCart

	Matriz de Confusão*	
Teste Negativo	438	62
Teste Positivo	114	154

*M 4.0 / C 0.98 / S 200

As árvores de decisão são funções que apresentam como dados de entrada um vetor de atributos e uma decisão como valor de saída. O funcionamento da árvore de decisão ocorre através da divisão de um conjunto de dados em subconjuntos de forma recursiva (GUARIGUATA et al, 2014).

O algoritmo SimpleCart se diferencia dos demais algoritmos devido ao tamanho da árvore, pois o número de folhas é menor em relação aos demais algoritmos, mostrando que esse método necessitou de um menor número de testes lógicos para a determinação das classes (BHAVSAR; GANATRA, 2016).

Entre as principais vantagens do uso das árvores de decisão, se destaca a fácil interpretação dos seus resultados, pois a classificação é obtida de forma explicita, simplificando a sua interpretação (MORAN et al., 2011).

CONSIDERAÇÕES FINAIS

O algoritmo *SimpleCart* se mostrou computacionalmente eficiente, pois os resultados foram gerados rapidamente, podendo desta forma poder vir a servir de subsídio para os gestores na tomada de decisão em saúde.

Entretanto, mesmo sendo caracterizado um modelo eficiente, e por gerar árvores pequenas, esse modelo não foi adequado para o banco de dados utilizado na pesquisa, apresentando o nível de Kappa considerado razoável, portanto podendo desta forma sere, realizados em estudos vindouros, a testagem deste método em outros bancos de dados.

REFERÊNCIAS

AHER, S. B.; LOBO, L., Comparative Study of Classification Algorithms. **International Journal of Information Technology**, 2012. **5**(2): p. 239-43.

AITKENHEAD, M. J. A co-evolving decision tree classification method. **Expert Systems with Applications**, v. 34, p. 18-25, 2008.

BEAGLEY, J.; GUARIGUATA, L.; WEIL, C.; MOTALA, A. A. Global estimates of undiagnosed diabetes in adults. **Diabetes Res Clin Pract**, 2014; 103(2):150-60.

BHAVSAR, H.; GANATRA, A. An Empirical Evaluation of Data Mining Classification Algorithms. **International Journal of Computer Science and Information Security**, v. 14, n. 5, p. 142, 2016.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. Belmont, **CA: Wadsworth International**, 1984. 93(99): p. 101.

DHAKATE, P., et al., Preprocessing and Classification in WEKA using different classifiers. **Journal of Engineering Research and Applications**. p. 2248-9622, 2014.

GUARIGUATA L, et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. **Diabetes Res Clin Pract**, 2014; 103(2): 137-49.

KALMEGH, S. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. **International Journal of Innovative Science, Engineering & Technology**, 2015. 2(2): p. 438-446.

KHARCHE, D.; RAJESWARI, K.; ABIN, D. Comparison of different datasets using various classification techniques with weka. **International Journal of Computer Science and Mobile Computing**, 2014. **3**(4): p. 389-393.

KOTZ, S.; JOHNSON, N. L. Encyclopedia of statistical sciences. **New York: John Wiley & Sons**; 1983. v.4, p.352-4.

SALLEY, K. E, et al. Glucose intolerance in polycystic ovary syndrome: a position statement of the Androgen Excess Society. **J Clin Endocrinol Metab**. 2007;92(12):4546-56.

SHAW, J. E.; SICREE, R. A.; ZIMMET, P. Z. Global estimates of the prevalence of diabetes for 2010 and 2030. **Diabetes Res Clin Pract**, 2010; 87(1): 4-14.

SMITH, J. W. et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. **In Proceedings of the Annual Symposium on Computer Application in Medical Care** (p. 261). American Medical Informatics Association, 1988.

SPEYBROECK, N. Classification and regression trees. **International journal of public health**, 2012. **57**(1): p. 243-246.

SRINIVASAN, B., MEKALA P. Mining Social Networking Data for Classification Using Reptree. **International Journal**, 2014. **2**(10).

STOVALL, D. W. et al. Assessment of insulin resistance and impaired glucose tolerance in lean women with polycystic ovary syndrome. **J Womens Health (Larchmt)**. 2011;20(1):37-43.

MORAN, L. J. et al. Diabetes risk score in the diagnostic categories of polycystic ovary syndrome. **Fertil Steril**. 2011;95(5):1742-8.

MOURA, E. C. et al. Research on chronic noncommunicable diseases in Brazil: meeting the challenges of epidemiologic transition. **Rev Panam Salud Publica**, 2012; 31(3): 240-5.

PORTER, R. S.; KAPLAN, J. L. **Manual Merck Diagnóstico e Tratamento**. 19ª ed. São Paulo: Roca; 2014.

TRUJILLANO, J. et al. Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. **Gaceta Sanitaria**, 2008, *22*(1), 65-72.