MÉTODOS DE CLASSIFICAÇÃO AUTOMÁTICA PARA PREDIÇÃO DO PERFIL CLÍNICO DE PACIENTES PORTADORES DO DIABETES MELLITUS

Glaucia Maria BRESSAN[1]
Beatriz Cristina Flamia de AZEVEDO[1]
Roberto Molina de SOUZA[1]

- RESUMO: O objetivo deste trabalho é estudar as relações entre os principais atributos que exercem influência no diagnóstico e controle do Diabetes Mellitus tipo 2 e gerar uma ferramenta de classificação automática que permita inferir sobre o índice glicêmico e que possa servir de auxílio na área médica para que o paciente portador do diabetes possa ser direcionado ao tratamento adequado. Os métodos propostos para tal tarefa são baseados no método de classificação Bayesiana, o qual utiliza o algoritmo BayesRule e é capaz de investigar incertezas probabilísticas nos dados, e no método de classificação por Árvores de Decisão, sendo esta uma ferramenta de classificação muito utilizada em mineração de dados devido à fácil interpretação dos resultados. Ambas as metodologias extraem regras de classificação linguística, o que permite a comparação de suas performances. De acordo com o processo de validação cruzada, o método de classificação Bayesiano com o algoritmo BayesRule apresenta 65% de acerto na tarefa de classificação para o grupo intervenção e 47,5% para o grupo controle. As árvores de Decisão com Poda apresentam 73,68% de acerto para o grupo intervenção e 69,23% para o grupo controle. Desta forma, os resultados obtidos neste trabalho se mostraram satisfatórios, podendo contribuir para controle e predição do desenvolvimento de pacientes portadores do Diabetes Mellitus tipo 2.

- PALAVRAS-CHAVE: Algoritmo BayesRule; árvores de decisão; redes bayesianas; índice glicêmico.

[1]Universidade Tecnológica Federal do Paraná - UTFPR, CEP: 86300-000, Cornélio Procópio, PR, Brasil, E-mail: *galbressan@gmail.com; beatrizflamiadx@hotmail.com; rmolina-souza@utfpr.edu.brr.*

1 Introdução

O número de casos diagnosticados de doenças crônicas cresce a cada ano, assim como o número de óbitos devido à descoberta tardia da doença e seus agravamentos (CASTANHO, 2013). O diagnóstico precoce e o prévio início do tratamento, são fatores determinantes para conter o desenvolvimento destas doenças, além de evitar complicações secundárias, norteando o gerenciamento de cuidados e hábitos do paciente.

De acordo com Whiting et al. (2011), o Diabetes Mellitus é uma das doenças crônicas mais comuns em quase todos os países e continua aumentando em número e significância. O desenvolvimento econômico e a urbanização impulsionam mudanças no estilo de vida, o que provoca a redução de atividades físicas e o aumento da obesidade (WHITING et al., 2011). Tais fatores exercem forte impacto no desenvolvimento do Diabetes Mellitus.

Segundo dados das Diretrizes da Sociedade Brasileira de Diabetes 2015-2016, uma epidemia de Diabetes Mellitus está em curso. Estima-se que a população mundial com diabetes seja da ordem de 387 milhões e que alcance 471 milhões em 2035 (MILECH et al., 2016). No Brasil, calcula-se que há mais de 15 milhões de pessoas portadoras da doença. Em números absolutos o Brasil ocupa o 4º lugar no mundo em número de diabéticos (15,4 milhões), apenas perdendo para: China (114,1 milhões), Índia (70,2 milhões) e Estados Unidos (32,8 milhões) (HUANG et al., 2017).

O diabetes é definido como um conjunto de doenças metabólicas que se desenvolve quando o pâncreas deixa de produzir insulina (hormônio que regula a concentração de açúcar (glicose) no sangue, glicemia) suficiente ou quando o organismo não é capaz de utilizar de forma eficaz a insulina que produz (OMDL, 2016). A ADA, *American Diabetes Association* (ADA et al., 2017), ressalta, basicamente, quatro classes clínicas de Diabetes: (i) *Diabetes mellitus tipo 1*, causado pela destruição das células beta-pancreáticas (β-pancreáticas) com consequente deficiência da produção da insulina e, consequentemente, na secreção deste hormônio; (ii) *Diabetes Mellitus tipo 2*, que se caracteriza por defeitos na secreção da insulina, mas, na maioria dos casos, na ação deste hormônio, estando grande parte dos pacientes com sobrepeso ou obesos, sendo estes fatores contribuintes para a diminuição da sensibilidade à insulina nos tecidos periféricos; (iii) *tipos específicos de anomalias genéticas das células* β-pancreáticas, classe associada a defeitos monogenéticos na função das células β-pancreática e, frequentemente, tem início em idade precoce; (iv) *Diabetes Mellitus Gestacional* definido como qualquer grau de intolerância à glicose, sendo primeiramente diagnosticada durante a gestação. Neste trabalho, o estudo é voltado para a predição do perfil clínico de pacientes portadores do Diabetes Mellitus tipo 2, de acordo com a variação de seu índice glicêmico, a fim de indicar o melhor tratamento para o paciente.

O Diabetes, seja qual for a classe, se não diagnosticado e tratado a tempo pode trazer complicações crônicas e agudas para o paciente, como por exemplo convulsões,

perda de consciência, danos aos olhos, ao coração, aos vasos sanguíneos e ao sistema nervoso, assim como o aumento do risco de doenças cardíacas e acidentes vasculares cerebrais. Nota-se que boa parte destas complicações podem levar a óbito; contudo, frequentemente, na declaração de óbito, não se menciona o Diabetes Mellitus como causa da morte, pelo fato de atribuírem às suas complicações (MILECH et al., 2016). Este fato contribui para que não seja possível mensurar exatamente o número de mortes causados pelo Diabetes. Entretanto, sabe-se que muitas dessas mortes poderiam ser evitadas com prévio conhecimento de controle glicêmico e mudança de estilo de vida (CASTANHO, 2013).

Neste trabalho, o objetivo é estudar as relações entre os principais atributos que podem influenciar no aumento ou redução do índice glicêmico, que é o indicador do Diabetes Mellitus tipo 2, e gerar uma ferramenta de classificação automática do perfil do paciente, de tal forma que o método de classificação seja de fácil interpretação na área médica, para indicação do tratamento adequado. Para isso, os atributos de entrada para os classificadores são selecionados a partir de indivíduos portadores do Diabetes Mellitus e a variável resposta (ou saída) do sistema de classificação é o aumento ou a redução do índice glicêmico. Desta forma, os resultados podem ser utilizados para inferir sobre novos pacientes com determinada precisão. Os classificadores utilizados para a tarefa de inferir sobre o aumento ou redução do indicador do Diabetes Mellitus tipo 2 são técnicas de aprendizado supervisionado: Redes Bayesianas, juntamente com o algoritmo BayesRule (HRUSCHKA et al., 2007), e Árvores de Decisão, com a utilização do algoritmo C4.5 (QUINLAN, 2014). A justificativa é dada por se tratarem de classificadores do tipo *White Box* (Caixa Branca), ou seja, revelam sua estrutura e proporcionam acesso à informação e às etapas de execução, a partir dos blocos de construção do algoritmo. Além disso, também proporcionam fácil interpretação dos resultados por meio de regras linguísticas, o que não ocorre com outros classificadores do tipo *Black Box* (Caixa Preta).

2 Materiais e métodos

Tanto as Redes Bayesianas, juntamente algoritmo BayesRule, quanto as Árvores de Decisão, permitem apresentar como resposta um conjunto de regras linguísticas do tipo "se-então", conforme apresentada no exemplo: "SE atributo w é *y1* E atributo x é *y2* ENTÃO a classe é z". Este tipo de resultado facilita a interpretação dos usuários do sistema, permitindo que pessoas sem conhecimento técnico avançado possam compreender os resultados obtidos. Além disso, ambos os métodos empregados neste trabalho concedem informações gráficas, o que auxilia na interpretação e análise do relacionamento dos atributos. A seguir são descritas as duas metodologias utilizadas.

2.1 Redes bayesianas e algoritmo BayesRule

O modelo de inferência Bayesiana consiste em analisar de maneira probabilística variáveis condicionadas de acordo com o *Teorema de Bayes*, descrito conforme Box e Tiao (BOX e TIAO, 2011): Suponha que $y' = (y_1, ...y_n)$ seja um vetor de n observações no qual a distribuição de probabilidade $p(y \mid \theta)$ dependa do valor de k parâmetros $\theta' = (\theta_1, ...\theta_k)$. Considere que θ tenha distribuição de probabilidade $p(\theta)$. Assim, dada a observação y, a probabilidade condicional de θ é representada pela equação (1), conhecida como *Teorema de Bayes*.

$$p(\theta \mid y) = \frac{p(y \mid \theta) \times p(\theta)}{p(y)} \tag{1}$$

No Teorema de Bayes, $p(\theta)$ é chamada de distribuição *a priori*, a qual fornece o que é sabido sobre θ sem o conhecimento dos dados e $p(\theta \mid y)$, conhecida como distribuição *a posteriori* de θ dado y, apresenta o que sabe-se sobre θ dado o conhecimento dos dados. Considerando que a informação sobre y é conhecida, $p(y \mid \theta)$ pode ser considerada uma função de θ, conhecida como *verossimilhança* de θ dado y. A verossimilhança é uma função que, por meio do conhecimento dos dados, modifica o conhecimento *a priori* de θ; portanto, representa a informação de θ proveniente dos dados. Em outras palavras, o Teorema de Bayes estabelece que a distribuição *a posteriori* de θ dado y é proporcional ao produto da distribuição *a priori* de θ dado o conhecimento dos dados pela verossimilhança de θ dado y. Ou seja:

$$posteriori \propto verossimilhanca \times priori \tag{2}$$

Na maioria dos sistemas reais, a variável de interesse é dependente ou está condicionada a várias outras. Mesmo podendo ser expandido para mais de uma variável, o Teorema de Bayes unicamente não é recomendado em casos envolvendo múltiplas variáveis, devido à complexidade dos cálculos (NEAPOLITAN et al., 2004). Diante dessa dificuldade, Pearl (1988) desenvolveu redes, as quais avaliam as interligações das variáveis por meio de suas estruturas, conhecidas como *Redes Bayesianas*.

As Redes Bayesianas podem ser definidas como um modelo gráfico probabilístico usado para representar o conhecimento sobre o domínio dos dados. Redes Bayesianas e classificadores Bayesianos aprendem relações de causa e consequência e podem combinar conhecimento *a priori* com padrões aprendidos à partir dos dados. Além disso, uma Rede Bayesiana permite que o usuário ou especialista interfira nos nós da rede e insira um conhecimento que se propaga nos outros nós, o que seria dificultoso em um sistema do tipo *Black Box*, como por exemplo, em uma Rede Neural Artificial.

Uma Rede Bayesiana é composta por uma estrutura de rede, que consiste em um grafo acíclico direcionado e um conjunto de tabelas de probabilidade. Os nós da estrutura da rede representam as variáveis e os arcos entre os nós representam relações de dependência entre as variáveis correspondentes. Um arco começando

em um nó A e terminando em um nó B estabelece A como pai de B e B como filho de A. Um exemplo de estrutura de rede pode ser visto na Figura 1, em que os nós em destaque representam os pais, os filhos e os pais dos filhos do nó A. Desta forma, uma rede Bayesiana pode ser utilizada como um classificador, calculando a probabilidade condicional de um nó, chamado nó classe, dados os valores das probabilidades dos outros nós.

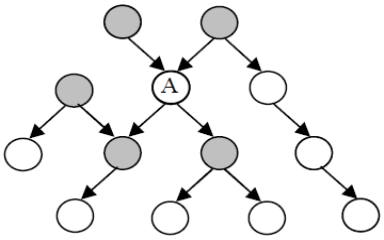

Figura 1 - Exemplo de uma estrutura de Rede Bayesiana.

Em Cheng et al. (2002), uma rede Bayesiana é representada por $BN =< N, A, \Theta >$, em que $< N, A >$ é um grafo acíclico direcionado em que cada nó $x_i \in N$ representa uma variável e cada arco $a \in A$ entre nós representa uma dependência probabilística entre os nós associados. Uma distribuição de probabilidade condicional θ_i é associada com cada nó $x_i \in N, i = 1, ..., n$, coletivamente representado por $\Theta = \theta_i$, que quantifica o quanto um nó depende de seus pais (CHENG et al., 2002). Conforme descrito em (PEARL, 2014), a independência condicional, que é a condição de Markov, permite o cálculo da distribuição conjunta de todas as variáveis, dado pela equação 3

$$P(x_1, x_2, ...x_n \mid BK) = \prod_{i=1}^{n} P(x_i \mid \pi_{xi}, BK), \tag{3}$$

em que BK representa o conhecimento antecedente, x_i é a i-ésima variável ou nó, e π_{xi} é o conjunto de pais de x_i. Portanto, uma Rede Bayesiana pode ser usada como uma representação do conhecimento que permite inferências.

Entretanto, apesar das Redes Bayesianas serem capazes de trabalhar com informações complexas, elas unicamente não são capazes de expressar conhecimento de forma suficientemente clara e semelhante ao raciocínio humano. Assim, um método chamado *BayesRule* (HRUSCHKA et al., 2007), usa o conceito de probabilidade máxima *a posteriori* para extrair um conjunto de regras probabilísticas do tipo "se-então". Esta combinação explora o potencial dos classificadores Bayesianos em relação à tarefa de classificação e o transforma em um conjunto de regras de classificação, tornando-o mais compreensível e favorecendo o entendimento das relações entre as variáveis e o processo de tomada de decisão.

O método *BayesRule* faz uso do algoritmo K2, o qual emprega a busca heurística para induzir uma Rede Bayesiana usando como entrada uma base de dados e uma lista ordenada de variáveis. Este algoritmo utiliza um método *greedy* para procurar a melhor estrutura de rede que represente os dados (HRUSCHKA et al., 2007). Além de utilizar a probabilidade máxima *a posteriori* o *BayesRule* emprega o conceito do Markov Blanket para reduzir o número e a complexidade de regras de classificação geradas pelo processo de extração. Isto permite a redução no tempo requerido para construir modelos quando mais de uma variável é designada como classe.

Seguindo esta aproximação, um classificador Bayesiano evidencia que o algoritmo de propagação deve ser usado para propagar os valores das variáveis e inferir o valor da variável classe. A confiança da regra pode ser definida utilizando resultados inferenciais. Com isso, a probabilidade dada à classe inferida pode ser usada como um valor de confiança e é embutido no algoritmo de inferência.

2.2 Árvores de decisão e o algoritmo C4.5

Árvores de Decisão constituem uma ferramenta de apoio à tomada de decisão que retorna como resultado um gráfico com estrutura semelhante a uma árvore invertida. Este método baseia-se em métodos e modelos estatísticos que utilizam aprendizado supervisionado para extrair conhecimento de um conjunto de dados e possibilitar inferências.

Comumente, na construção de uma Árvore de Decisão, utiliza-se um algoritmo do tipo *TDIDT (Top-Down Induction Decision Tree)*, o qual induz uma estrutura de Árvore de Decisão dividindo o conjunto de dados em subgrupos cada vez mais uniformes, baseando-se no método dividir e conquistar (LANTZ, 2013). Este processo de divisão termina quando o subconjunto apresentar apenas uma classe de dados ou quando não for possível melhores resultados. Além disso, o processo de divisão pode ser interrompido por critérios preestabelecidos (CICHOSZ, 2015).

Uma árvore de decisão consiste em um conjunto de dados particionado em grupos conhecidos como *nós*. O *nó* do topo é chamado de *nó raiz*, este é selecionado usando métricas de seleção próprias de cada algoritmo. Abaixo do *nó raiz*, localizam-se os *nós internos*, que constituem os *ramos* da árvore. Ao final de cada *ramo* estão as folhas, que são nós terminais, e que representam uma classe de dados para as regras da Árvore de Decisão, que são do tipo "se-então".

Pode-se dizer que o processo de modelagem de uma Árvore de Decisão se desenvolve em duas fases: a fase de construção, em que ocorre a partição do conjunto de dados e a estrutura da árvore é construída e a fase de poda da árvore, a qual pode ser considerada um método de otimização da Árvore de Decisão.

A fase de construção é um processo recursivo em que se estabelece a estrutura da Árvore de Decisão de acordo com um critério de partição. Existem diferentes critérios que podem ser empregados para se construir uma Árvore de Decisão. Cada algoritmo é composto por critérios de divisão e poda distintos, o que, consequentemente, gera Árvores de Decisão distintas. Neste trabalho, foi escolhido o algoritmo C4.5 por ser um dos algoritmos mais frequentes e atuais na literatura.

O algoritmo C4.5 é uma evolução do algoritmo ID3 (*Iterative Dichotomiser*) (QUINLAN, 1996).

O algoritmo C4.5 utiliza a *Taxa de Ganho* como critério de partição dos dados; por este critério, quando o número de instâncias for menor que um determinado valor, a divisão do conjunto de dados é interrompida (ROKACH e MAIMON, 2008).

Para se introduzir os conceitos da Taxa de Ganho, primeiramente é necessário apresentar os princípios de *Entropia* e *Ganho de Informação*. A Entropia mede o grau de pureza dos dados, indicando o quão homogêneo ou heterogêneo estão as classes de dados. Esta medida varia de 0 a 1, no qual 0 indica classes mais homogêneas e 1 classes mais heterogêneas (LANTZ, 2013). Matematicamente, a Entropia pode ser expressa por 4.

$$Entropia(y, S) = \sum_{c_j \in dom(y)} -\frac{|\sigma_{y=c_j} S|}{|S|} \times log_2 \frac{|\sigma_{y=c_j} S|}{|S|} \tag{4}$$

O ganho de informação usa a Entropia com medida de pureza, conforme apresentado em (5). O algoritmo decide qual atributo deve se dividir e usa a Entropia para calcular a mudança de homogeneidade, resultante da divisão de cada atributo possível.

$$GanhoInf(a_i, S) = Entropia(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \times Entropia(y, \sigma_{a_i=v_{i,j}} S) \tag{5}$$

Finalmente, a Taxa de Ganho, expressa por (6), representa a razão do ganho de informação pela Entropia.

$$TaxadeGanho(a_i, S) = \frac{GanhoInf(a_i, S)}{Entropia(a_i, S)} \tag{6}$$

A Taxa de Ganho tende a favorecer atributos que apresentem denominador muito pequenos. Primeiro, o Ganho de Informação é calculado para todos os atributos. Então, o atributo que obter a melhor Taxa de Ganho é selecionado, como consequência são considerados atributos que apresentaram performance ao menos semelhante a média do Ganho de Informação (ROKACH e MAIMON, 2008). Em outras palavras, o critério de seleção estipulado pela Taxa de Ganho, seleciona um teste para maximizar a taxa (6), sujeita a restrição de que o Ganho de Informação deve ser pelo menos tão grande quanto o ganho médio de todos os testes examinados.

A segunda fase da modelagem por Árvores de Decisão é chamada de poda. Esta fase é responsável por reduzir a complexidade da árvore, otimizando o modelo em construção. A complexidade diz respeito ao número de ramos e folhas que a árvore apresenta; quanto menor a árvore, menor a complexidade e maior a facilidade de compreensão dos resultados gerados. Árvores de Decisão com alta complexidade apresentam problemas de ruídos e *overffiting*, os quais reduzem a eficiência e acurácia dos modelos. *Overfitting* ocorre quando a Árvore de Decisão

perde a capacidade de generalização para dados desconhecidos durante o processo de modelagem. Ao aumentar o número de nós, o erro de treinamento geralmente diminui, enquanto em algum momento o erro de generalização se torna pior (ROKACH e MAIMON, 2008).

O processo de poda utiliza medidas estatísticas para remover os ramos que apresentam menor confiabilidade (PATIL et al., 2010). Consequentemente, este processo otimiza a operação computacional, eliminando o *overfitting* e melhorando a classificação para dados desconhecidos. O processo de poda também varia de acordo com o algoritmo. *Cost-Complexity, Reduced Error Pruning, Minimum Error Pruning, Error Based Pruning*, são alguns dos critérios que podem ser usados para podar árvores (ROKACH e MAIMON, 2008).

Alguns critérios de poda eliminam parte das árvores, chamadas de sub-árvores, e, então, as reposicionam como folhas. O critério de poda usado pelo C4.5 denomina-se *Poda Baseada em Erro*, uma evolução de um outro critério conhecido como *Poda Pessimista*.

Rokach e Maimon (2008) descrevem a Poda Baseada em Erro é descrita da seguinte forma: a taxa de erro é estimada usando o limite superior do intervalo de confiança estatístico para proporções, conforme (7):

$$\varepsilon_{UB}(T,S) = \varepsilon(T,S) + Z_\alpha \times \sqrt{\frac{\varepsilon(T,S) \times (1-\varepsilon(T,S))}{|S|}} \tag{7}$$

em que $\varepsilon(T,S)$ indica a taxa de classificação errada da árvore T com relação ao conjunto de treinamento S; Z é o inverso da distribuição cumulativa normal padrão; e α é o nível de significância desejado. A subárvore(T,t) denota a subárvore enraizada no nó t, e $maxfilho(T,t)$ indica o filho mais frequente para o nó t. S_t corresponde a todas as instâncias de S que estão ao alcance de t. O procedimento varre todos os nós de baixo para cima e compara os seguintes valores:

I) $\varepsilon_{UB}(subarvore(T,t), S_t)$;

II) $\varepsilon_{UB}(poda(subarvore(T,t),t), S_t)$;

III) $\varepsilon_{UB}(subarvore(T, maxfilho(T,t)), Smaxfilho(T,t))$.

De acordo com o valor mais baixo, o procedimento deixa a árvore como está, ou poda o nó t, ou ainda substitui t pela árvore enraizada pelo $maxfilho(T,t)$, (ROKACH; MAIMON, 2008).

3 Estudo de caso

O conjunto de dados utilizado neste trabalho é proveniente de uma pesquisa clínica, com delineamento de coorte prospectivo, do tipo anterior-posterior, com duração de 12 meses, divido em 3 avaliações: inicial, após 6 meses e após 12 meses. A pesquisa foi realizada no município de Bandeirantes, Paraná, conforme descrita em Castanho (2013). Os dados foram coletados a partir de 77 pacientes, portadores de Diabetes Mellitus, divididos em dois grupos: 39 pacientes pertencem ao grupo *intervenção* (Grupo 1) e 38 pacientes pertencem ao grupo *controle* (Grupo

2), sendo que a intervenção é baseada no uso do método de gerenciamento de caso de enfermagem, com atendimento multiprofissional, ações educativas em grupos, cuidado domiciliar e abordagem telefônica (CASTANHO, 2013). O grupo controle consiste no tratamento convencional da doença, com uso de medicamentos.

De acordo com Castanho (2013) e considerando a correlação entre estas variáveis e seu impacto no desenvolvimento da doença, os principais atributos que podem influenciar na variação dos indicadores do Diabetes Mellitus tipo 2 são: Idade, Nível de Triglicérides e o Tempo de Evolução do diagnóstico da doença. Este último trata-se do período de tempo, contabilizado em anos, referido pelo paciente, desde o início da doença até o momento do estudo.

Assim, esses três atributos foram considerados como atributos de entrada nos modelos de classificação por Redes Bayesianas. Além dos três atributos citados anteriormente, o Índice de Massa Corporal (IMC), a Circunferência Abdominal, a Renda per Capita e o Nível de Escolaridade também exercem forte influência no processo de inferência do diagnóstico doença, de acordo com a expertise do especialista. Desta forma, para os modelos de classificação por Árvores de Decisão foram considerados este total de 7 atributos citados.

A Tabela 1 apresenta as medidas resumo (média e desvio padrão) para os atributos utilizados nos sistemas de classificação. Esta tabela é elaborada de acordo com os grupos em estudo (intervenção e controle) e com a saída do sistema (aumento ou redução do índice glicêmico). Os atributos Idade, Tempo de Evolução da doença e Renda per Capita apresentam efeito cruzado na redução ou aumento do índice glicêmico, segundo os grupos. Por exemplo, no grupo intervenção, as pessoas mais jovens, em média, apresentaram redução, enquanto no grupo controle, foram as mais velhas. Isto sugere que a intervenção não necessariamente é o melhor encaminhamento para todos os indivíduos. Desta forma, a proposta de classificação deste trabalho, baseada nas características do indivíduo, é relevante para direcionar os indivíduos ao tratamento mais adequado.

4 Modelos propostos de classificação

Selecionados os atributos de entrada dos sistemas de classificação, nesta seção são descritas as propostas de classificação de acordo com as metodologias empregadas neste trabalho: a classificação por Redes Bayesianas e a classificação por Árvores de Decisão.

4.1 Classificação por redes bayesianas

O modelo de classificação utilizando Redes Bayesianas é elaborado com auxílio computacional do software livre GENIE [1]. São selecionados 3 atributos para que seja possível aplicar o Método BayesRule para extração de regras linguísticas de uma rede Bayesiana, considerando a correlação entre tais atributos e seu impacto no

[1]Desenvolvido pela Universidade de Pittsburgh (https://dslpitt.org/genie/)

Tabela 1 - Medidas resumo dos atributos de entrada dos sistemas de classificação

Atributo	Grupo	Saída	N	Média	Desvio Padrão
Idade (anos)	Intervenção	Aumentou	7	51,29	6,50
		Reduziu	31	49,71	6,64
	Controle	Aumentou	13	47,00	10,52
		Reduziu	26	51,81	5,32
Triglicérides (mg/dL)	Intervenção	Aumentou	7	180,86	98,06
		Reduziu	31	192,32	103,00
	Controle	Aumentou	13	157,77	97,93
		Reduziu	26	173,38	104,50
Tempo de Evolução (Anos)	Intervenção	Aumentou	7	8,71	5,12
		Reduziu	31	8,21	8,45
	Controle	Aumentou	13	5,06	5,06
		Reduziu	26	7,27	5,44
IMC (Kg/m^2)	Intervenção	Aumentou	7	31,21	2,26
		Reduziu	31	27,92	5,83
	Controle	Aumentou	13	34,33	10,65
		Reduziu	26	31,17	5,61
Circunferência Abdominal (cm)	Intervenção	Aumentou	7	104,93	3,27
		Reduziu	31	97,18	12,06
	Controle	Aumentou	13	107,69	16,46
		Reduziu	26	102,94	11,84
Renda per capita (R$)	Intervenção	Aumentou	7	340,95	243,21
		Reduziu	31	304,44	194,76
	Controle	Aumentou	13	241,31	189,72
		Reduziu	26	375,91	210,90
Escolaridade (anos)	Intervenção	Aumentou	7	4,57	3,36
		Reduziu	31	4,97	3,94
	Controle	Aumentou	13	6,15	5,24
		Reduziu	26	5,85	4,83

desenvolvimento da doença. De acordo com os testes, os atributos que apresentam relações significativas e expressam os melhores resultados sobre o aumento ou a diminuição do indicador do Diabetes Mellitus tipo 2 são: idade, nível de triglicérides e o tempo de evolução do diagnóstico da doença.

Os atributos de entrada, foram discretizados no software GENIE em intervalos de classes linguísticos, de acordo com a massa de dados. Os atributos "idade" e "nível de triglicérides" foram discretizados em 4 classes (ou categorias) e o atributo "tempo de evolução" em 3 classes, devido a distribuição dos dados.

Conforme mencionado na Seção 3, o conjunto de dados em estudo é dividido em dois grupos, intervenção (Grupo 1) e controle (Grupo 2). Assim, para cada grupo é elaborada uma Rede Bayesiana a partir da aplicação dos algoritmos *K2* e *Greedy Thick Thinning* (LERNER e MALKA, 2011). As estruturas das Redes Bayesianas obtidas são apresentadas na Figura 2, respectivamente, para os Grupos 1 e 2. O nó "Saída" representa o índice glicêmico e os demais nós representam os atributos de entrada selecionados. Os arcos, por sua vez, denotam os relacionamentos entre os atributos e a saída do classificador.

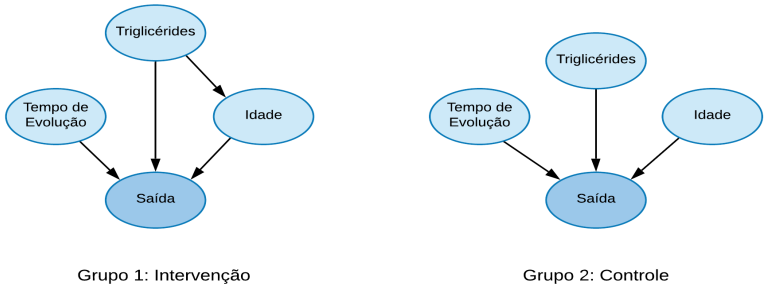

Figura 2 - Estrutura das Redes Bayesianas Geradas.

Após a estrutura das redes, o algoritmo BayesRule é utilizado para extrair o conjunto de regras linguísticas. De acordo com o de número de atributos e a quantidade de classes discretizadas, é possível obter 48 combinações linguísticas, as quais originam as regras do tipo "se-então" de cada modelo de classificação. Como exemplo, uma das regras de classificação da rede do grupo intervenção pode ser descrita da seguinte forma: "Se *triglicérides* é classe 1 e *idade* é classe 4 e *tempo de evolução* é classe 1 então o índice glicêmico *reduziu* com probabilidade 0,625".

Para validação destes modelos de classificação, é utilizado o conceito de validação cruzada (ROKACH e MAIMON, 2008), em que o conjunto de dados é aleatoriamente dividido em n subconjuntos mutualmente exclusivos para teste. A cada verificação de acerto da classificação, um destes subconjuntos é considerado como teste e os demais como treinamento, e assim sucessivamente, até que todos os subconjuntos sejam testados. Neste caso, foi considerado $n = 10$, e por esta razão, o conjunto de dados é então dividido em 4 conjuntos de treinamento e teste, tanto para o Grupo 1 (que contém 39 instâncias) quanto para o Grupo 2 (que contém 38 instâncias). Desta forma, a validação cruzada para o Grupo 1 é executada com 3

subconjuntos de 10 instâncias e um subconjunto com 9 instâncias para teste e para o Grupo 2 é executada com 3 subconjuntos de 10 instâncias e um subconjunto com 8 instâncias para teste. Cada vez que um subconjunto de teste é selecionado, todos os demais dados são utilizados para o treinamento.

Com estes modelos de classificação foram obtidos os índices de acerto na classificação de 65% para o grupo intervenção e 47,5% para o grupo controle.

4.2 Classificação por árvores de decisão

Para a elaboração dos modelos de Árvores de Decisão são considerados mais atributos de entrada, pois, apenas os 3 principais utilizados anteriormente são insuficientes para o algoritmo determinar as relações de dependência entre eles. Desta forma, novamente considerando a correlação entre os atributos e seu impacto no desenvolvimento da doença, de acordo com os testes realizados, os atributos que apresentam relações significativas e expressam os melhores resultados sobre o aumento ou a diminuição do indicador do Diabetes Mellitus tipo 2 são 7: idade, nível de triglicérides, o tempo de evolução do diagnóstico da doença, índice de massa corporal, circunferência abdominal, renda per capita e nível de escolaridade.

Para elaboração dos modelos foi utilizado o software WEKA[2] por ser um software livre e possuir o algoritmo C4.5 implementado.

Também para os grupos de intervenção e controle, as Árvores de Decisão são geradas para a tarefa de classificação visando a predição do perfil clínico de pacientes portadores do indicador do Diabetes Mellitus tipo 2, considerando a variação do índice glicêmico. Contudo, para os modelos de Árvores de Decisão, não é necessário que os dados estejam na forma categórica, já que o algoritmo permite o uso de variáveis numéricas, o que diminui a perda de informação dos dados.

Visando a comparação de resultados dos classificadores, primeiramente são obtidas as estruturas de Árvores de Decisão sem a utilização da técnica de poda, conforme apresentadas na Figura 3. Desta forma, para o grupo intervenção, foi gerada uma árvore contendo 7 folhas e tamanho 13, com índice de acerto na classificação igual a 73,68%, utilizando validação cruzada como anteriormente. Para o grupo controle, a árvore apresentou 4 folhas, tamanho 7 e 64,10% de acerto na classificação.

Os modelos de Árvores de Decisão gerados com implementação da Poda Baseada em Erro estão apresentados na Figura 4. Estes modelos apresentaram 5 folhas, tamanho 9 e 73,68% de acerto no modelo de grupo intervenção. Para a estrutura do grupo controle, não houve mudança na estrutura gráfica da árvore, no entanto, o nível de acerto aumentou para 69,23%, de acordo com a validação cruzada considerando também $n = 10$, para fins de comparação.

Analisando a estrutura de cada Árvore de Decisão gerada, é possível extrair as regras linguísticas "se-então" de acordo com o número de folhas da estrutura modelada. Cada ramo da árvore do Grupo 1 e do Grupo 2 gera uma regra linguística. Desta forma, um exemplo de regra linguística extraída da árvore podada

[2](Waikato Environment of Knowledge Analysis - http://www.cs.waikato.ac.nz)

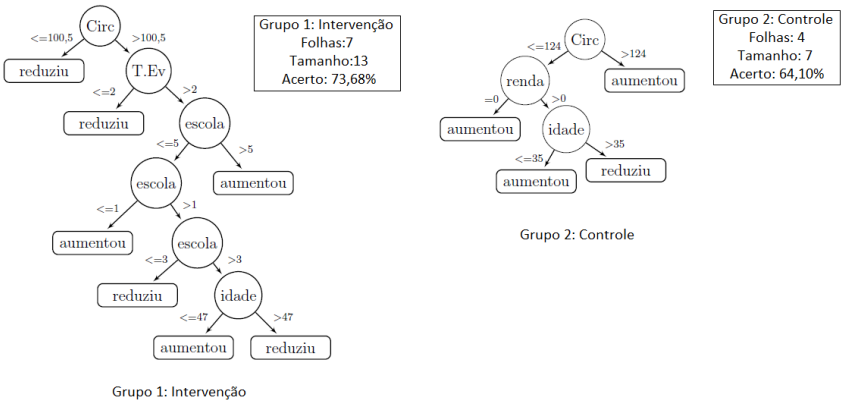

Figura 3 - Árvores de Decisão sem Poda.

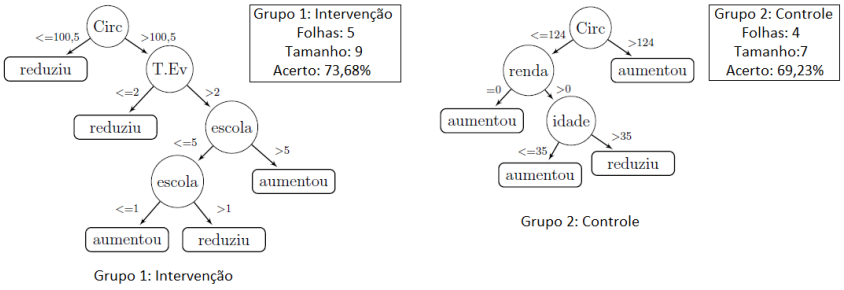

Figura 4 - Árvores de Decisão com Poda.

do grupo intervenção pode ser descrita da seguinte maneira: "Se *circunferência abdominal* é maior que 100,5 cm e *tempo de evolução* é maior que 2 dias e *escolaridade* é maior que 5 anos então o índice glicêmico *aumentou*". E um exemplo de regra extraída da árvore podada do grupo controle é: "Se *circunferência abdominal* é maior que 124 cm, então o índice glicêmico *aumentou*".

Conclusões

Neste trabalho, são apresentadas duas propostas de classificação automática para predizer o perfil clínico de pacientes portadores do Diabetes Mellitus tipo 2, considerando a variação do seu índice glicêmico: utilizando Redes Bayesianas, juntamente com o algoritmo BayesRule para extração de regras, e utilizando Árvores de Decisão. Estas duas metodologias foram escolhidas por serem abordagens do tipo *White Box*, apresentando resultados de fácil compreensão, podendo ser utilizados por profissionais da área da saúde, sem a necessidade de conhecimentos técnicos avançados de modelagem e análises de dados.

Observando-se os resultados obtidos, os modelos de classificação gerados por Árvores de Decisão apresentaram melhores índices de acerto no processo de validação cruzada. E apesar de serem modelos compostos por apenas sete atributos de entrada, é possível perceber a importância da técnica de poda para redução da complexidade e otimização dos modelos, aumentando o índice de acerto. Nas estruturas de Árvores de Decisão do grupo intervenção pode-se notar graficamente a redução do número de folhas e tamanho das estrutura, sem que o índice de acerto seja prejudicado. Já nos modelos do grupo controle, apesar de não haver mudança gráfica das estruturas após o processo de poda, houve alteração no índice de acerto; isto pode ser justificado pelo fato de que a validação cruzada utiliza conjuntos de teste e treinamento distintos a cada processamento, até que todo o conjunto de dados seja testado.

Na prática, um profissional da saúde, a partir dos atributos medidos de seu paciente, pode direcioná-lo da seguinte maneira: aplica-se a base de regras gerada no grupo controle; caso haja redução no índice glicêmico (saída da regra), aloca-se o paciente para este grupo. Caso não haja redução, aplica-se a base de regras gerada no grupo intervenção e, havendo redução, aloca-se o paciente para este grupo. Não havendo redução em nenhum dos grupos, pode-se alocar o paciente no grupo controle, por questões de custo.

As duas ferramentas de classificação adotadas geram uma base de regras linguísticas de classificação, o que contribui significativamente para a interpretação da classificação pelo usuário, pois a leitura das regras geradas não exige que o usuário tenha conhecimento técnico sobre os algoritmos utilizados, além de permitir o acesso à informação contida nos nós das Redes Bayesianas ou nos ramos das Árvores de Decisão (métodos *White Box*). Por esta razão, pode-se dizer que ambas as metodologias podem ser empregadas para inferir sobre novos dados de pacientes, ou seja, a partir dos atributos medidos de um paciente, pode-se indicá-lo ao tratamento adequado, direcionando-o ou não à intervenção, para que haja maior sucesso na

redução dos indicadores do Diabetes Mellitus tipo 2. Por se tratar de um estudo biológico, em que existe uma variabilidade intrínseca aos sujeitos, os resultados podem ser considerados válidos. Como perspectivas futuras, pretende-se elaborar um modelo de classificação *Fuzzy* que seja capaz de descrever as fronteiras de forma a considerar sobreposições e incluir outras categorias na classe de saída que representa o índice glicêmico.

Agradecimentos

Os autores agradecem a Fundação Araucária pela bolsa de estudos concedida e aos dois revisores e editores pelas sugestões e comentários.

BRESSAN, G. M.; AZEVEDO, B. C. F.; SOUZA, R. M. Automatic classification methods for predicting clinical profile of patients with diabetes mellitus. *Rev. Bras. Biom.,* Lavras, v.38, n.2, p.257-273, 2020.

- *ABSTRACT: The goal of this paper is to study the relationships between the main attributes that influence the diagnosis and control of Diabetes Mellitus Type 2 and to generate an automatic classification tool that allows inferring about the glycemic index and which can be used as a medical aid in order to the patient with diabetes can be directed to the appropriate treatment. The methods proposed for this task are based on Bayesian Classification method, which uses the BayesRule algorithm and is able to investigate probabilistic uncertainties in the data, and on the classification method using Decision Trees, which is a classification tool widely used in data mining due to easy interpretation of the results. Both methodologies extract linguistic classification rules, which allows the comparison of their performances. According to the cross-validation process, the Bayesian classification method with the BayesRule algorithm presents 65% accuracy in the classification task for the intervention group and 47.5% for the control group. The Pruning Decision Trees present 73.68% accuracy for the intervention group and 69.23% for the control group. Then the results obtained in this study are satisfactory, and may contribute to the control and prediction of the development of patients with Diabetes Mellitus Type 2.*

- *KEYWORDS: BayesRule algorithm; decision trees; Bayesian networks; glycemic index.*

Referências

ADA et al. *Standards of medical care in diabetes 2017.* American Diabetes Association. Diabetes care, Am Diabetes Assoc, v.40, 2017.

BOX, G. E.; TIAO, G. C. *Bayesian inference in statistical analysis.* New York: John Wiley & Sons, 2011.

CASTANHO, R. M. *Efeito do uso do método de gerenciamento de caso sobre o controle glicêmico de pessoas com diabetes mellitus tipo 2*. Tese (Doutorado). Universidade Federal do Paraná, Curitiba PR, 2013.

CHENG, J. et al. Learning bayesian networks from data: an information-theory based approach. *Artificial intelligence*, v.137, n.1−2, p.43-90, 2002.

CICHOSZ, P. *Data mining algorithms: explained using R*. Malden: Wiley Online Library, 2015.

HRUSCHKA, E. R. et al. Markov-blanket based strategy for translating a bayesian classifier into a reduced set of classification rules. In: IEEE. *7th International Conference on Hybrid Intelligent Systems (HIS 2007)*. Kaiserlautern, Germany, 2007. p.192-197.

HUANG, Y. et al. *IDF diabetes atlas estimates for the global diabetes prevalence of adults aged 18 to 99 years*. 2017. A446-A446 p.

LANTZ, B. *Machine learning with R*. Birmingham: Packt Publishing, 2013.

LERNER, B.; MALKA, R. Investigation of the k2 algorithm in learning bayesian network classifiers. *Applied Artificial Intelligence*, v.25, n.1, p.74-96, 2011.

MILECH, A. et al. *Diretrizes da Sociedade Brasileira de Diabetes (2015-2016)*, São Paulo: AC Farmacêutica, 2016.

NEAPOLITAN, R. E. et al. *Learning bayesian networks*. Upper Saddle River: Pearson Prentice Hall, 2004.

OMDL, S. *Informe mundial sobre la diabetes*. Organización Mundial de la Salud: Ginebra, n.4, 2016.

PATIL, D. D.; WADHAI, V. M.; GOLHALE, J. A. Evaluation of decision tree pruning algorithms for complexity and classification accuracy. *International Journal of Computer Applications*, v.11, n.2, p.23-30, 2010.

PEARL, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Los Angeles: Elsevier, 2014.

QUINLAN, J. R. Improved use of continuous attributes in C4.5. J*Journal of Artificial Intelligence Research*, v.4, p.77-90, 1996.

QUINLAN, J. R. C. *Programs for machine learning*. San Francisco: Elsevier, 2014

ROKACH, L.; MAIMON, O. *Data mining with decision trees: Theory and applications*. Singapore: World Scientific, 2008.

WHITING, D. R. et al. IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*, v.94, n.3, p.311-321, 2011.

Recebido em 15.07.2019.

Aprovado após revisão em 01.10.2019.