rasbran
Revista da Associação Brasileira de Nutrição

Artigo Original

Árvores de Decisão como Método de Mineração de Dados: Análise de Prontuários de uma Clínica Escola de Nutrição

Decision Trees as a Data Mining: Analysis of Medical Records of the Nutrition Clinical School

Ademir Roberto Freddo[1], Márcia Fernandes Nishiyama[1], Kesia Zanuzo[1], Eloá Angélica Koehnlein[1], Rafaela Cristina da Silva Ramos[1]

[1] *Universidade Federal da Fronteira Sul*

Endereço para correspondência: Ademir Roberto Freddo - ademir.freddo@uffs.edu.br

Palavras-chave

Análise de dados
Aprendizado de Máquina
Registros Médicos
Identificação de doenças

Este artigo descreve a aplicação de árvores de decisão como método de mineração de dados nos prontuários da Clínica Escola de Nutrição da Universidade Federal da Fronteira Sul (UFSS), campus Realeza. Árvores de decisão é uma técnica de aprendizado de máquina, utilizada para reconhecimento de padrões em inteligência artificial na análise de dados. Para que se possa aplicar esta técnica, se faz necessária a existência de uma massa de dados com diversos atributos significativos. Portanto, a Clínica Escola de Nutrição da UFFS, campus Realeza, possui uma grande quantidade de prontuários que foram utilizados na mineração de dados. Utilizou-se 1339 prontuários. Destes, foram selecionados apenas atributos ou variáveis que estavam preenchidos em todos os prontuários a fim de não comprometer os resultados e criação dos modelos. Assim aplicou-se a técnica de árvores de decisão aos prontuários para identificar como classe final as respectivas doenças: dislipidemia, diabetes e hipertensão. Portanto, para cada doença, a partir dos prontuários selecionados, foi criada uma árvore de decisão ou modelo. Este possui um conjunto de atributos ou variáveis mais significativas para a identificação e caracterização da doença. Além disso, a confiabilidade ou eficiência da árvore na identificação da doença foi definida com base nos 30% dos 1339 prontuários. Portanto, 402 registros foram utilizados para testar os modelos, sendo para dislipidemia a eficiência foi de aproximadamente 89%. Na hipertensão e diabetes obteve-se uma eficiência de aproximadamente 94%. Isto significa, o quanto as regras criadas a partir das árvores de decisão são eficientes na identificação das doenças. Estas regras podem ser utilizadas para o desenvolvimento de alguma ferramenta computacional para auxílio no diagnóstico e também para prever e classificar doenças.

Keywords

Data analysis
Machine Learning
Medical Records
Disease prevention

This article presents the application of decision trees as a method of data mining in the medical records of the Nutrition Clinical School Federal University of the South Frontier (UFFS) in Realeza city . Decision trees is a machine learning technique, used to recognize patterns in artificial intelligence in data analysis. In order to apply decision trees, it is necessary to have data with several important attributes. Therefore, the Nutrition Clinical School, has a large number of medical records that were used in data mining. We used 1339 medical records. Of these, we selected only attributes or variables that were filled in all the medical records to create of the models. From these 1339 medical records, the decision tree technique was applied to identify the final diseases as dyslipidemia, diabetes and hypertension. Thus, for each disease, a tree or model was created with a set of attributes or variables more significant for the identification and characterization of the disease. In addition, the reliability or efficiency of the tree in the identification of the disease was defined based on 30% of 1339 medical records. Therefore, 402 records were used to test the models, and for dyslipidemia the efficiency was approximately 89%. In hypertension and diabetes an efficiency of approximately 94% was obtained. This means, how much the rules created from the decision trees are efficient in the identification of the diseases. These rules can be used to develop some computational tool to aid in diagnosis and also to predict and classify diseases.

INTRODUÇÃO

A Clínica Escola de Nutrição da Universidade Federal da Fronteira Sul (UFFS), campus Realeza, realiza atendimentos à comunidade nas áreas de avaliação e diagnóstico nutricional, educação nutricional, reeducação alimentar e acompanhamento ambulatorial de indivíduos e grupos específicos da população. Na realização das consultas, os estudantes com acompanhamento dos professores nutricionistas, fazem as anotações no prontuário do paciente. Estas anotações são utilizadas pelo estudante e professor nutricionista para definição de diagnósticos, acompanhamento, bem como orientações nutricionais. Porém, com base nos dados presentes no prontuário, é difícil para o nutricionista correlacionar dados do paciente, identificar padrões, prever futuros problemas de saúde e descobrir potenciais riscos à saúde. Além da análise individual, o nutricionista também não consegue realizar uma comparação entre históricos de pacientes a fim de analisar semelhanças e padrões entre os prontuários.

Os prontuários, mais especificamente, os dados, são uma fonte de conhecimento que podem ser utilizados no ensino na área nutricional, na descoberta de padrões, na tomada de decisões no diagnóstico e futuras avaliações clínicas do paciente, bem como na prevenção e combate às doenças.

A interpretação e análise dos dados pode ser feita por meios estatísticos. Entretanto, há outra técnica, chamada de mineração de dados[1], que permite a descoberta de modelos ou padrões. A mineração de dados utiliza técnicas de aprendizado de máquina (*i.e machine learning*), entre elas, árvores de decisão[2].

O aprendizado de máquina é um ramo da inteligência artificial que tem como objetivo desenvolver técnicas capazes de ensinar ao computador a aprender a partir das próprias experiências. A aplicação prática do aprendizado de máquina inclui processamento de linguagem natural, diagnósticos médicos, processamento de imagens (visão computacional), reconhecimento de padrões (em escrita e fala), robótica computacional, entre outros[3].

Para que se possa aplicar o aprendizado de máquina, se faz necessária a existência de uma massa de dados de treinamento e teste, com diversos atributos ou variáveis significativas. Em outras palavras, para que o computador possa aprender ele necessita dos dados e das respostas[4]. Logo, a clínica escola de nutrição da UFFS, campus Realeza, possui uma quantidade de dados, mais precisamente, prontuários que podem ser utilizados no aprendizado de máquina.

Portanto, o trabalho de análise dos dados deste projeto visou descobrir os padrões nos prontuários da clínica escola nutrição, aplicando a técnica de mineração de dados denominada de árvores de decisão[5]. A partir dos padrões ou modelos criados, foi possível identificar quais são os fatores, atributos ou características mais importantes na identificação e prevenção da dislipidemia, diabetes e hipertensão.

METODOLOGIA DO TRABALHO

O método utilizado para extrair conhecimento a partir dos prontuários (*Knowledge Discovery in Databases* - KDD), utilizou seguintes etapas[6]: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação.

Utilizou-se 1339 prontuários como dados de entrada. Destes, foram selecionados apenas atributos ou variáveis que estavam preenchidos em todos os prontuários a fim de não comprometer os resultados e criação dos modelos. A partir destes 1339 prontuários, aplicou-se a técnica de árvores de decisão para identificar como classe final as respectivas doenças: dislipidemia, diabetes e hipertensão. Assim, para cada doença, a partir dos 1339 prontuários, foi criada uma árvore de decisão ou modelo que correspondem a regras, baseadas em atributos, para identificar as doenças escolhidas. Estas regras podem ser utilizadas para o desenvolvimento de alguma ferramenta computacional para auxílio no diagnóstico, prevenção e classificação de doenças.

Etapa 1: Seleção de dados

Os dados utilizados na mineração de dados foram obtidos a partir dos prontuários dos pacientes atendidos na Clínica Escola de Nutrição. A utilização dos dados foi autorizada pelo Comitê de Ética em Pesquisa (CEP) da UFFS, no parecer 980.593 de 19 de março de 2015.

Atualmente, os prontuários são registrados em planilhas eletrônicas. Os dados selecionados referem-se a aproximadamente 2000 pacientes registrados (*i.e* 2000 prontuários eletrônicos) com as variáveis ou atributos apresentados no Quadro 1. Todos os atributos descritos no Quadro 1 foram utilizados para a mineração de dados.

Quadro 1: variáveis ou atributos dos prontuários utilizados na mineração de dados.

Atributo/Variável	**Descrição (valor preenchido)**
IDADE	Em anos
SEXO	Masculino ou Feminino
PROFISSÃO	Principal ocupação professional
CONSUMO BEBIDAS	Se consome bebidas alcoólicas (SIM/NÃO)
ANT_FAM (4 POSSIBILIDADES)	Antecedentes familiares (algum tipo de doença)

Quadro 2 (continuação)

Atributo/Variável	Descrição (valor preenchido)
ATIVIDADE FISICA	Se pratica atividades físicas (SIM/NÃO)
HORAS DE SONO	Número de horas que dorme por dia
ANSIEDADE_ALIMENTACAO	Se possui ansiedade durante as refeições (SIM/NÃO)
CONSUMO DE LÍQUIDOS	Se consome líquidos durante as refeições (SIM/NÃO)
CONSOME ASSISTINDO TV	Se durante as refeições assiste televisão (SIM/NÃO)
QUANTIDADE DE ARROZ	Regularmente / raramente/ nunca
QUANTIDADE DE LEGUMES	Regularmente / raramente/ nunca
REFRIGERANTE/SUCOS ARTIFICIAIS	Regularmente / raramente/ nunca
CONSUMO DE EMBUTIDOS	Regularmente / raramente/ nunca
TEMPEROS PRONTOS E CONDIMENTOS	Regularmente / raramente/ nunca
CONSUMO DIÁRIO DE VERDURAS E LEGUMES	Quantas horas por dia
CONSUMO DIÁRIO DE FRUTAS	Quantas horas por dia
CONSUMO DE ÁGUA/DIA	Em litros
IMC	Índice de Massa Corporal
INTOLERÂNCIA A LACTOSE	SIM/NÃO
CIRROSE	SIM/NÃO
TABAGISMO	Se pratica (SIM/NÃO)
ALERGIAS	SIM/NÃO
DIABETES	SIM/NÃO
DISLIPIDEMIA	SIM/NÃO
HIPERFERRITINEMIA	SIM/NÃO
HIPERTENSÃO	SIM/NÃO

Etapa 2: Pré-processamento

No pré-processamento ocorreu a identificação de dados duplicados, incorretos, faltantes, valores atípicos e discrepantes. Assim, alguns prontuários foram retirados da base de dados por não apresentarem os valores dos atributos (Quadro 1) preenchidos. Com isto, dos 2000, foram selecionados 1339 prontuários. A retirada dos prontuários com atributos não preenchidos é importante para não comprometer a execução da técnica de árvore de decisão e consequente geração do modelo ou padrão. Assim apenas os prontuários que possuíam todos os atributos com valores preenchidos foram mantidos no conjunto de dados.

Etapa 3: Preparação dos dados

A preparação dos dados consiste na discretização (*e.g.* categorização de valores numéricos em intervalos de valores) dos atributos contínuos que forem necessários para a classificação, categorização de atributos, padronização, normalização e escalonamento dos dados. Nesta fase, também realizou-se a conversão dos dados para o formato ARFF (*Atribute-Relation File Format*) exigido pela ferramenta WEKA (*Waikato Environment for Knowledge Analysis*)[7] utilizada para a mineração de dados.

Etapa 4: Mineração de dados

Na mineração de dados aplicou-se o método de aprendizagem de máquina denominado de árvores de decisão, utilizando-se da ferramenta computacional WEKA[7]. Nesta fase é criado o modelo ou padrão que consiste no desenho de uma árvore com o os atributos ou variáveis observadas a partir dos prontuários. Todos os 1339 prontuários foram submetidos a etapa de mineração de dados. O objetivo foi criar 3 modelos ou árvores de decisão para as respectivas doenças: dislipidemia, diabetes e hipertensão. Cada árvore de decisão representa um conjunto de regras para identificação das doenças citadas.

Árvore de decisão é um método para aprendizagem indutiva. Uma árvore de decisão é construída de cima para baixo (*top-down*), usando o princípio de dividir para conquistar. Um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema[8].

Uma árvore de decisão possui nós, ramos, folhas e percursos. A Figura 1 ilustra um exemplo de representação de uma árvore de decisão para classificação de frutas. Na Figura 1, os nós representam cor, tamanho, forma e sabor. Cada nó contém um teste de um atributo. Os ramos correspondem a um possível valor dos atributos partindo de um nó correspondente. Na Figura 1, verde, amarelo, médio, grande e doce são exemplos de ramos e possíveis valores de atributos. As classes são representadas por folhas e indicam uma classificação. Maçã, limão, banana são exemplos de classes na Figura 1. O percurso, em uma árvore de decisão, é formado do nó raiz até a folha e representa uma regra de classificação.

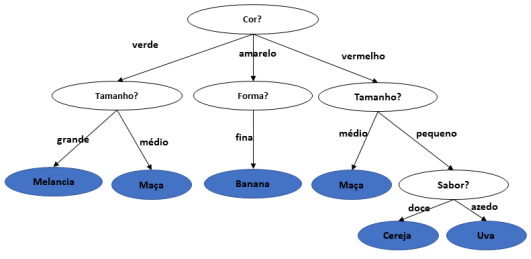

Figura 1: Exemplo de uma árvore de decisão para classificação de frutas.

Uma árvore de decisão também pode ser transformada em um conjunto de regras (se-então). Utilizando a Figura 1 pode-se criar a seguinte regra: SE (cor = verde) E (tamanho = grande) ENTÃO Fruta = Melancia.

A classificação de uma instância em uma árvore de decisão é feita inicialmente pelo nó raiz, testando o atributo especificado por este nó. Em seguida, move-se para baixo na árvore, seguindo o caminho determinado pelas respostas aos testes dos nós até atingir alguma folha. Há vários algoritmos para aprendizagem em árvores de decisão, entre eles: ID3[8], C4.5[5,9] e C5.0[10,11]. Neste trabalho utilizou o algoritmo C4.5[5].

O algoritmo C4.5 constrói uma árvore de decisão de cima para baixo verificando primeiro qual é o melhor atributo para a raiz da árvore. A seleção do melhor atributo consiste em um teste estatístico para determinar quão bem o atributo sozinho classifica os exemplos de treinamento. A partir do nó raiz, os descendentes são criados para cada valor possível deste atributo, ou seja, estende-se a árvore adicionando um ramo para cada valor do atributo. Esse processo é repetido usando os exemplos de treinamento para selecionar o melhor atributo seguinte. O processo continua até que a árvore classifique todos os exemplos de treinamento e todos os atributos tenham sido utilizados[8]. A seleção do melhor atributo utiliza conceitos como a entropia e ganho[4].

A entropia é a medida da impureza ou o grau de desordem de um conjunto de dados de treinamento. Utiliza-se a seguinte fórmula para a entropia:

$$Entropia(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Onde:

S: um conjunto de dados contendo exemplos positivos e negativos de algum conceito alvo;

p+: proporção de exemplos positivos em S;

p-: proporção de exemplos negativos em S.

A entropia é zero quando todos os membros de S são positivos, portanto pertencem a mesma classe. A entropia contém valor um quando há um número igual de exemplos positivos e negativos. O valor entre zero e um corresponde a números diferentes de exemplos positivos e negativos. Se um atributo pode assumir c valores diferentes então a entropia de S relativa a essa classificação é definida como:

$$Entropia(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Onde: **pi** é a proporção de S pertencendo a classe i.

A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia. Após a definição de entropia como uma medida da impureza em uma coleção de exemplos de treinamento, pode-se definir agora a medida da efetividade de um atributo para classificar os dados de treinamento. Para isso utiliza-se a medida chamada ganho de informação, que é simplesmente a redução esperada na entropia causada pelo particionamento dos exemplos por este atributo. Mais precisamente, o ganho de informação de um atributo A, relativo a uma coleção de exemplos S, é definido como[8]:

$$Ganho(S, A) = Entropia(S) - \sum_{v \in Valores(A)} \left|\frac{S_v}{S}\right| Entropia(S_v)$$

Onde:

Valores(A) é o conjunto de todos possíveis valores para atributo A;

Sv é o subconjunto de S para qual o atributo A tem valor **v**.

O ganho de informação é a redução esperada da entropia, sendo uma medida quantitativa que mede quão bem um dado atributo separa os exemplos de treinamento de acordo com a classificação alvo, ou seja, qual é a habilidade de um atributo em discriminar as classes.

Etapa 5: Extração do conhecimento

Ao aplicar a técnica de árvore de decisão aos 1339 registros/prontuários, foram criados os modelos ou padrões que foram interpretados nesta etapa para geração ou extração de conhecimento. A seguir nos resultados a interpretação dos resultados obtidos, mais precisamente das árvores de decisão geradas.

RESULTADOS

Nesta seção são descritos os resultados obtidos pela aplicação da técnica de mineração de dados (árvores de decisão) para a base de dados com 1339 registros/prontuários. A técnica foi aplicada para as seguintes categorias ou classes: dislipidemia, hipertensão e diabetes. A seguir a descrição dos modelos ou árvores geradas para cada doença.

Árvore de Decisão para Dislipidemia

A Figura 2 ilustra a árvore de decisão para a dislipidemia. De acordo com a árvore gerada, os atributos relevantes para identificação da doença são, hiperferritinemia, hipertensão, IMC, idade e a presença ou não de diabetes. Assim, de todos os atributos (Quadro 1) preenchidos no prontuário, apenas os apresentados na Figura 2 foram considerados importantes para a prevenção e detecção da doença. Com a geração da árvore é possível criar regras, como por exemplo: SE (possui hiperferritinemia) E (possui hipertensão) E (idade superior a 73 anos) E (IMC acima de 32.24) E (não possui diabetes) ENTÃO pode desenvolver a doença de dislipidemia. O final de

cada ramo da árvore (sim/não) indica ou não a possibilidade da doença com base na construção das regras.

A confiabilidade das regras é gerada pelo sistema WEKA. Neste caso, a técnica classificou corretamente 89% dos registros pela identificação da dislipidemia. Portanto, em 89% dos casos, as regras se aplicadas nos prontuários, indicarão ou não a presença da dislipidemia.

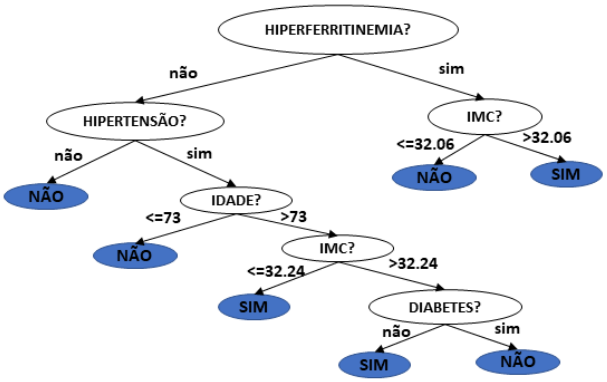

Figura 2: Árvore de Decisão gerada para Dislipidemia.

Árvore de Decisão para Hipertensão

A Figura 3 ilustra a árvore de decisão gerada para a doença hipertensão. Ao final de cada ramo da árvore, há um retângulo com a informação sim ou não. Isto corresponde a possibilidade ou não da hipertensão. Para chegar ao final de cada ramo da árvore, há uma série de atributos que o algoritmo considerou significativos. Neste caso, apenas os atributos idade, consumo de água, IMC e a existência ou não de diabetes foram considerados importantes. Por exemplo, observa-se que a existência de diabetes, com idade superior a 49 anos e IMC superior a 26,98 kg/m2c, classificam um paciente com a probabilidade de ter hipertensão. Porém, neste mesmo ramo, se o IMC for inferior a 26,98, o paciente não possui risco de hipertensão.

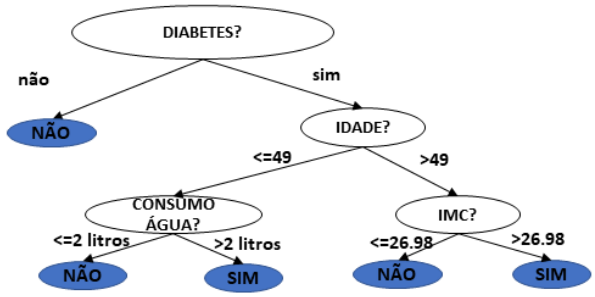

Figura 3: Árvore de Decisão gerada para Hipertensão.

Comparando a quantidade de registros da Tabela 1 com os registros da árvore, observa-se que a árvore possui poucos atributos. Isto significa que os outros atributos não foram relevantes para a criação do modelo. A confiabilidade para a classificação de um paciente com hipertensão, gerada pela árvore de decisão foi de 94%. Portanto, a técnica ou modelo obtido consegue classificar corretamente 94% dos registros, o que corresponde a 1267 registros.

Árvore de Decisão para Diabetes

A Figura 4 ilustra a árvore de decisão gerada para a doença diabetes na disposição de regras se (condição) então/senão. De acordo com a Figura 4 há vários ramos da árvore que são gerados principalmente pelos atributos hipertensão, hiperferritinemia, idade, sexo, consumo de bebidas, consumo de verduras e dislipidemia. Neste caso, o experimento gerou mais ramos do que os experimentos anteriores (dislipidemia e hipertensão). Portanto foram considerados mais atributos e geradas mais regras. A árvore gerada classifica corretamente 94% dos registros para a doença diabetes. Mesmo assim, o atributo mais importante na árvore é relacionado a existência de hipertensão. Por exemplo, caso o paciente tenha hipertensão, com idade maior que 49 anos, tenha dislipidemia, seu IMC <= 35.8, consumo diário de verduras inferior a duas porções e não ingere bebida alcoólica, pode ter a predisposição de desenvolver diabetes. Assim, ao acompanhar a sequencia se/então/senão, gera-se várias regras.

```
Se (HIPERTENSÃO = sim)
    Então Se (IDADE > 49)
                Então Se (DISLIPIDEMIA = sim)
                            Então Se (IMC > 35.8)
                                        Então NÃO DIABETES
                                        Senão Se (CONSUMO_DIARIO_VERDURAS >  2) (porções)
                                                    Então NÃO DIABETES
                                                    Senão Se (CONSUMO_BEBIDAS = sim)
                                                                Então NÃO DIABETES
                                                                Senão SIM DIABETES
                            Senão Se (SEXO = masculino)
                                        Então Se (CONSUMO_BEBIDAS = sim)
                                                    Então NÃO DIABETES
                                                    Senão Se (CONSUMO_DIÁRIO_VERDURAS > 3) porções
                                                                Então NÃO DIABETES
                                                                Senão SIM DIABETES
                                        Senão Se (IDADE > 65)
                                                    Então  Se  (CONSUMO_DIÁRIO  DE  FRUTAS  >3)  porções
                                                                Então NÃO DIABETES
                                                                Senão SIM DIABETES
                                                    Senão NÃO DIABETES
                Senão Se (HIPERFERRITINEMIA =sim)
                            Então SIM DIABETES
                            Senão NÃO DIABETES

Se (HIPERTENSÃO = não)
    Então Se (IDADE > 64)
                Então Se IMC > 21.5
                            Então NÃO DIABETES
    Senão  NÃO DIABETES
```

Figura 4: Árvore de Decisão gerada para Diabetes.

DISCUSSÃO

A utilização da mineração de dados para selecionar informações importantes na área da saúde, pode auxiliar na prevenção de doenças[12] e no diagnóstico médico[13].

Neste trabalho utilizou-se a mineração de dados obtidos de avaliações nutricionais para identificação de doenças, sendo que, cada doença corresponde a uma classe final denominada dislipidemia, hipertensão ou diabetes.

Os modelos gerados, mais precisamente, as árvores de decisão, podem ser utilizadas para o desenvolvimento de alguma ferramenta computacional para auxílio no diagnóstico, prevenção e classificação de doenças.

Nas árvores geradas, ao caminhar pelos ramos, são criadas regras que descrevem os atributos ou características mais significativas para prevenção e identificação das doenças. Portanto, os modelos gerados podem auxiliar a tomada de decisões no atendimento da clínica escola, bem como na identificação de padrões não detectados até o momento nos dados armazenados dos pacientes atendidos.

Além de identificar e prever doenças supracitadas, a mineração de dados pode também ajudar a diferenciar os diagnósticos médicos[14]. Um exemplo é a utilização de técnicas de mineração de dados para diferenciar o diagnóstico médico na obstrução biliar por cálculo ou por câncer[15]. Outro exemplo é utilização dos dados médicos para prevenção de doenças cardíacas[16]. Como neste trabalho, outros[17,18] procuram identificar a diabetes. Em[17] utiliza-se a correlação estatística de dados[17]. Em[18] utiliza-se árvores de decisão para identificar diabetes, porém como dados de entrada, prescrições médicas. Outros trabalhos[19,20] também procuram identificar o diabetes, porém utilizando outras técnicas de aprendizagem de máquina como redes neurais e máquinas de vetores de suporte (*Support Vector Machines* – SVMs)

Neste trabalho, os resultados da árvore de decisão para os 1339 prontuários, indicaram que muitos dos atributos, presentes na base de dados, são irrelevantes na produção dos modelos, portanto, seu cadastro na Clínica Escola de Nutrição da UFFS, campus Realeza, não é importante para a identificação de alguma doença. Esta conclusão refere-se apenas aos dados presentes na base de dados utilizada. Outras bases de dados podem resultar em outras conclusões e consequente identificação de outros atributos irrelevantes. Portanto, o valor dos atributos influencia na produção do modelo. Assim, trabalhos que utilizam técnicas de arvores de decisão, podem gerar diferentes modelos mesmo utilizando atributos ou características semelhantes.

Os dados utilizados em mineração de dados podem ser extraídos, como neste trabalho, de banco de dados que armazenam prontuários. Porém pode-se utilizar mineração de dados para diagnóstico médico a partir de características extraídas de imagens médicas[21].

Enfim, há vários estudos com a utilização de mineração de dados para identificação, diagnóstico e prevenção de doenças. O que diferem os estudos são as técnicas utilizadas de inteligência artificial, o objetivo ou classe final que deseja identificar, bem como os dados de entrada constituídos de atributos e seus respectivos valores, sendo prontuários ou até características extraídas de imagens médicas.

CONCLUSÕES

Este artigo descreveu os resultados da utilização da mineração de dados com o objetivo de identificar ou prever doenças a partir dos dados de prontuários nutricionais de pacientes da Clínica Escola de Nutrição da UFFS, campus Realeza.

Aos dados, foi aplicada mineração de dados, mais precisamente a técnica de aprendizagem de máquina denominada de árvores de decisão. Esta técnica gerou um modelo que corresponde uma árvore de decisão onde cada ramo apresenta uma regra para identificação ou não das doenças dislipidemia, hipertensão e diabetes.

A confiabilidade dos modelos, mais precisamente, o quanto os dados/atributos podem ser eficazes na identificação das doenças, foi verificada pela ferramenta (WEKA) que gerou as árvores. Isto indica a confiabilidade das regras. Os resultados foram razoavelmente satisfatórios para hipertensão e diabetes obtendo aproximadamente 94% de confiabilidade nos modelos para diabetes e hipertensão. Para a dislipidemia os resultados não foram satisfatórios, pois obteve-se uma confiabilidade de aproximadamente 89%.

Para a melhoria dos resultados, como trabalhos futuros, sugere-se a necessidade de utilizar maiores quantidades de entradas, realizar combinações de atributos, identificar a relevância dos atributos, bem como utilizar outras técnicas de mineração de dados.

REFERÊNCIAS

1. Witten IH, Frank E. Data Mining–Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers; 2005.
2. Berry MJA, Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Support. New York: Wiley Computer Publishing. 1997.
3. Russel S,N. Artificial Intelligence: A Modern Approach. 2nd ed.: Campus; 2004.
4. Mitchell TM. Machine Learning: McGraw–Hill Science/Engineering/Math; 1997.

5. Quinlan JR. C4.5 Programs for Machine Learning San Diego California: Morgan Kaufmann Publishers; 1993.
6. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996: p. 17.
7. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, I.H. W. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 2009.
8. Duda RHPSD. Pattern Classification. 2nd ed.: Willey Interscience; 2002.
9. Quinlan JR. Induction of Decision Trees. Machine Learning. 1986; 1.
10. Quinlan JR. Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research. 1996; 4.
11. Quinlan JR. Is C5.0 Better Than C4.5?. [Online].; 1997 [cited 2018 Maio 10. Available from: http://rulequest.com/see5-comparison.html.
12. Vijiyarani S. Sudha S. Disease Prediction in Data Mining Technique – A Survey. International Journal of Computer Applications & Information Technology. 2013.
13. Lavrac N. Keravnou E. Zupan B. Intelligent data analysis in medicine. Encyclopedia of computer science and technology, 42(9), 113-157, 2000.
14. Collazos K. Barreto JM. Pellegrini GF. Analise do Prontuário médico para a utilização com KDD. Congresso Brasileiro de Informática em Saúde–CBIS, 2000.
15. Steiner MTA. Soma NY. Shimizu T. Nievola JC. Lopes F. Smiderle A. Data-Mining como Suporte a Tomada de Decisões - uma Aplicação no Diagnostico Medico. XXXVI Simpósio Brasileiro de Pesquisa Operacional. 23, 96-107, 2004.
16. Soni J. Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications. New York. v. 17, n. 8, p. 43-48, 2011.
17. Marinov M. Data mining technologies for diabetes: a systematic review. Journal of Diabetes Science and Technology, Thousand Oaks, v. 5, n. 6, p. 1549-1556, 2011.
18. Toussi M. Lamy J. Le Toumelin P. Venot A. Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. BMC Med. Informat. Decis. Making 9–28, 2009.
19. Pham HNA. Triantaphyllou E. Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization. Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803, 2008.
20. Sapna S. Tamilarasi A. Data mining–Fuzzy Neural Genetic Algorithm in predicting diabetes. Department Of Computer Applications (MCA), K.S.R College of Engineering, Research Journal on Computer Engineering, March, 2008.
21. Costa AF. Traina AJM. Mineração de Imagens Médicas Utilizando Características de Forma. 2012.

Submissão: 26/08/2018
Aprovado para publicação: 07/09/2019