Métodos Computacionais para Predição de Doenças Cardíacas

Sene Cassamá
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID: 000-0002-8056-2489

Vitória de A. Xavier
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID: 0000-0001-5731-1908

Thifany Ketuli S. de Souza
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0000-0002-0138-1285

José Filipe S. de Andrade
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID:0000-0002-9346-102

Wellington. P. dos Santos
Departamento de Engenharia Biomédica
Universidade Federal de Pernambuco
Recife, Brasil
ORCID: 0000-0003-2558-6602

Resumo — As doenças cardíacas constituem as principais causas de morte em todo o mundo. O que pode ser muito difícil para os profissionais da área de saúde fazerem as previsões dos ataques cardíacos que acontecem por ano. Existem muitas informações ocultas no setor de saúde que podem ser úteis no momento da tomada de decisões. Os algoritmos de mineração de dados como *Naive Bayes, J48, MLP, Random Forest, Random Tree, SVM* foram aplicados neste trabalho para analisar a base de dados que prevê doenças cardíacas. Com a mineração de dados o setor de saúde consegue prever os conjuntos de padrões de dados e consegue interpretá-los com uma boa precisão o que conduzirá a bons resultados para diminuir as doenças cardíacas. Este trabalho tem como objetivo analisar de forma crítica uma base de dados existente sobre as doenças cardíacas e propor as sugestões sobre os resultados obtidos.

Palavras - Chave — Algoritmo, Classificação, Doenças Cardíacas, Métodos Computacionais.

Abstract – Heart disease is the leading cause of death worldwide. This can be very difficult for health professionals to predict heart attacks that happen every year. There is a lot of information hidden in the health sector that can be useful when making decisions. Data mining algorithms such as Naive Bayes, J48, MLP, Random Forest, Random Tree, SVM have been applied in this work to analyze the database that provides heart disease. With data mining the health sector can predict the sets of data patterns and can interpret them with good accuracy which will lead to good results to decrease heart disease. This paper aims to critically analyze an existing database on heart disease and propose the suggestions.

Keywords: Algorithm, Classification, Heart Disease, Computational Methods

I. INTRODUÇÃO

É um fato conhecido em todo o mundo que o coração é um órgão vital para o funcionamento do corpo humano. Se este órgão é afetado, isso influencia todas as partes vitais do corpo. Segundo a Organização Mundial de Saúde (OMS), estima-se que ocorrem anualmente 17,5 milhões de mortes em todo mundo por causa das doenças cardíacas. A OMS estima que até 2030 cerca de 23,7 milhões de pessoas irão morrer devido a uma doença cardíaca, fato proporcional à qualidade de vida das pessoas [1].

As doenças cardíacas envolvem problemas funcionais no coração, tais como infecção de músculos, coração irregular, problemas no ritmo, anomalias nas válvulas cardíacas, etc. Isso pode contribuir para insuficiência cardíaca, podendo gerar ataques cardíacos ou derrames [2].

Investir na detecção precoce de doenças cardíacas é fundamental para evitar grandes gastos de recursos públicos e privados destinados para o tratamento dessas patologias [3]. Dessa forma, a utilização de aprendizagem de máquina desempenha um papel significativo para o auxílio no diagnóstico em saúde, visto que o método de classificação permite a identificação de anormalidades através da observação e a conexão de padrões presente em um banco de dados cuja associação com as classes é inteligível ao homem [4]. Ademais, é fato que a aquisição de bons atributos dos pacientes contribui para uma base de dados mais direcional quanto a resolução do problema. Neste estudo, a base de dados passou por métodos computacionais a fim da busca de classificadores com métricas satisfatórias.

II. MATERIAIS E MÉTODOS

O banco de dados utilizado para a realização da pesquisa foi o *Heart Disease UCI* que foi obtido em formato ARFF presente no *Kaggle*, que é uma comunidade online de pesquisadores de ciência de dados, e fornecido pela disciplina de Inteligência Artificial Aplicada a Engenharia Biomédica do Departamento de Engenharia Biomédica da UFPE. O arquivo completo possui 76 atributos, porém o do presente trabalho possui 14 atributos, sendo um deles a classe do problema, e 303 instâncias [5,6].

Esta base de dados tem como classe binária o target sendo a situação “0” a ausência de doença e a situação “1” a presença de alguma anormalidade. São 138 instâncias classificadas com ausência de doença e 165 instâncias com presença de alguma doença. O arquivo não diz qual patologia cada instância apresenta, apenas dita a presença ou a ausência de qualquer doença cardíaca.

Os atributos adquiridos de cada paciente são: *age*, idade em anos; *sex*, sexo binário (1 para masculino ou 0 para feminino); cp, o tipo de dor torácica, sendo 0 angina típica, 1 angina atípica, 2 dor não anginosa ou 3 assintomática; trestbts, a pressão arterial em repouso em mmHg; *chol*, colesterol sérico em mg/dl; fbs, medida do açúcar no sangue em jejum, sendo 1 para > 120 mg/dl e 0 para o oposto; restecg, resultados eletrocardiográficos em repouso, sendo 0 normal, 1 com anormalidade da onda ST-T ou 2 mostrando hipetrofia ventricular esquerda; *thalac*, frequência cardíaca máxima alcançada; *exang*, presença(1) ou ausência(0) de

angina induzida por exercícios; oldpeak, depressão de ST induzida por exercícios em relação ao repouso; slope, inclinação do segmento ST de pico do exercício, sendo 0 inclinação ascendente, 1 inclinação plana ou 2 inclinação descendente; ca, número de vasos principais, de 0 à 3, coloridos por fluroscopia; *thal*, a talassemia, sendo 0 normal, 1 o defeito corrigido e 2 o defeito reversível [7].

Para avaliar os classificadores aplicados neste banco de dados, foi utilizado o software de mineração de dados WEKA, tendo por parâmetros padrões a opção de teste de *cross-validation* com 10 *folds* e realizando 30 repetições. Em busca de bons resultados nas métricas das classificações, utilizou-se o banco de dados mencionado para um primeiro resultado e fez-se uma seleção de atributos do mesmo arquivo para um segundo resultado. A seleção de atributos foi feita no WEKA com os algoritmos PSO, genético e colônia de formigas, onde todos forneceram os mesmos atributos para a nova base de dados: cp, *thalach*, *exang*, *oldpeak*, *slope*, ca e *thal*.

Em todos os dois bancos de dados utilizou-se os seguintes classificadores: MultiLayer Perceptron (MLP), com taxa de aprendizado de 0.3 e 500 iterações com uma camada de 50 neurônios, uma de 100 neurônios e duas camadas de 50; *Support Vector Machine (SVM)*, com parâmetro de folga 0.01, 0.1 e 1.0, para o *Polynomial Kernel linear*, com grau 2, com grau 3 e *kernel RBF; Naive Bayes; Bayes Net; J48; Random Tree; Random Forest* com 10, 50 e 100 árvores.

Nos testes foram calculadas as métricas de interesse como a acurácia, sensibilidade, especificidade, área da curva ROC e índice kappa. A acurácia consiste, basicamente, no grau de proximidade de uma estimativa com seu parâmetro ou valor verdadeiro [8]. As métricas de sensibilidade e especificidade, por outro lado, são utilizadas na descrição da precisão de um resultado em um mesmo teste; sendo a sensibilidade correspondente ao percentual de resultados positivos verdadeiros e a especificidade ao percentual de negativos verdadeiros (em condições ideais, tanto especificidade quanto sensibilidade apresentam valores iguais a 100%). A área da curva ROC avalia a capacidade discriminativa de um teste, sendo uma métrica de qualidade geral que avalia se o experimento é capaz de classificar corretamente (assim como a sensibilidade e especificidade, quanto mais próximo de 100% os resultados, mais próximo do ideal) [9]. Por fim, o índice kappa que mede o grau de concordância das avaliações feitas nas mesmas amostras por diversos avaliadores, podendo variar de -1 até +1, onde quanto maior o valor de kappa, mais forte a concordância [10]. A partir desses valores foi possível verificar a eficácia do método na classificação da base.

Além disso, outros trabalhos com procedimentos de diagnóstico de doenças baseados em técnicas computacionais foram analisados, como a doença de Alzheimer [11, 12], o mal de Parkinson [13] e câncer de mama [14, 15] foram estudados, e procurou-se, portanto, verificar a aplicabilidade dessa metodologia para predição de doenças cardíacas. Com isso, observou-se que os resultados foram promissores nesses trabalhos e que o método poderia oferecer uma aplicabilidade relevante na predição de doenças cardíacas, os resultados mostraram isso após aplicação desse método.

III. Resultados e Discussões

A. Classificação Utilizando a Base Integral

Na Tabela 1 estão indicados os índices adotados para cada um dos seis classificadores com os melhores resultados observados. A Tabela 2 apresenta os valores de acurácia e índice *kappa*, enquanto que na Tabela 3 estão os valores de sensibilidade, especificidade e curva ROC dos classificadores referenciados na tabela anterior.

TABELA 1: Índices e os respectivos classificadores

Índice	Classificador
1	BayesNet
2	Naive Bayes
3	SVM Kernel Linear (c =1,0)
4	SVM Kernel Polinomial de Grau 2 (c = 0,1)
5	SVM Kernel Polinomial de Grau 3 (c = 0,01)
6	SVM Kernel Polinomial de Grau 3 (c = 0,1)

TABELA 2: Acurácia e índice kappa das melhores classificações obtidas

Classificador	Acurácia (%)	Kappa
1	82,899 ± 6,783	0,654 ± 0,137
2	82,750 ± 6,674	0,649 ± 0,136
3	83,010 ± 6,920	0,652 ± 0,135
4	82,877 ± 6,467	0,649 ± 0,133
5	82,382 ± 6,256	0,638 ± 0,130
6	82,319 ± 6,675	0,639 ± 0,137

TABELA 3: Sensibilidade, Especificidade e Curva ROC das 6 melhores classificações obtidas

Classificador	Sensibilidade	Especificidade	Curva ROC
1	0,789 ± 0,106	0,862 ±0,106	0,908 ± 0,054
2	0,762 ± 0,115	0,881 ±0,115	0,894 ± 0,058
3	0,728 ± 0,116	0,915 ±0,075	0,822 ± 0,068
4	0,729 ± 0,114	0,912 ±0,076	0,820±0,067

5	0,704 ± 0,115	0,924 ±0,070	0,814 ± 0,065
6	0,745 ± 0,116	0,889 ±0,089	0,817 ± 0,068

.

B. *Classificação Após Seleção de Atributos*

Assim como nos testes sem seleção de atributos foram calculadas as mesmas métricas de interesse (acurácia, índice *kappa*, sensibilidade, especificidade e área da curva ROC) após a seleção dos atributos. Os melhores classificadores foram os mesmos presentes na Tabela 1. Por isso, foi mantida a ordenação dos índices adotados.

Portanto, na Tabela 4 são mostrados os valores de acurácia e índice kappa dos classificadores, assim como na Tabela 5 estão presentes os valores de sensibilidade, especificidade e área da curva ROC.

TABELA 4: Acurácia e Índice Kappa após seleções

Classificador	Acurácia (%)	Kappa
1	83,143 ± 6,974	0,654 ± 0,137
2	82,750 ± 6,674	0,649 ± 0,136
3	83,010 ± 6,920	0,652 ± 0,135
4	82,877 ± 6,467	0,649 ± 0,133
5	82,382 ± 6,256	0,638 ± 0,130
6	82,319 ± 6,675	0,639 ± 0,137

TABELA 5: Sensibilidade, Especificidade e Curva ROC após seleção

Classificador	Sensibilidade	Especificidade	Curva ROC
1	0,781 ± 0,115	0,873 ± 0,091	0,901 ± 0,058
2	0,765 ± 0,113	0,879 ± 0,080	0,880 ± 0,067
3	0,739 ± 0,114	0,898 ± 0,074	0,819 ± 0,065
4	0,704 ± 0,116	0,915 ± 0,072	0,809±0,068
5	0,622 ± 0,129	0,912 ± 0,070	0,767 ± 0,075
6	0,703 ± 0,114	0,917 ± 0,071	0,810 ± 0,064

C. *Comparação de classificadores com e sem seleção de atributos*

Após obtidos os resultados, torna-se necessário fazer uma comparação entre os principais valores de referência para estimar qual o método mais adequado para a classificação do problema.

Na Figura 1 e 2 há a comparação entre as acurácias dos 6 melhores classificadores antes e depois da seleção de atributos adotada.

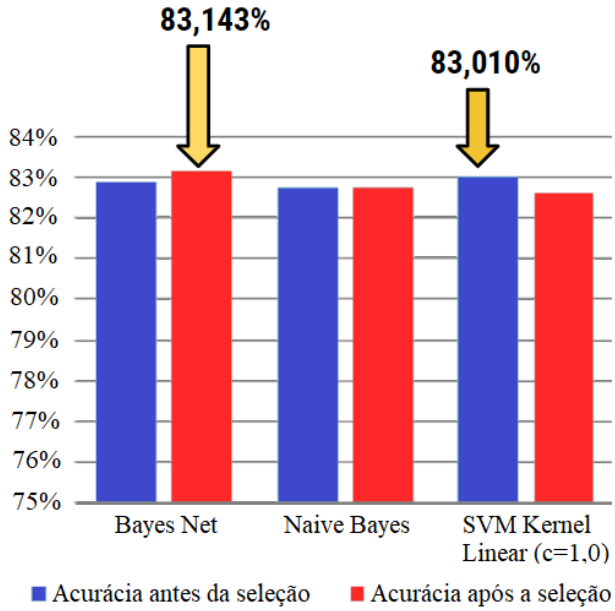

Fig. 1. Comparação entre acurácias antes e depois da seleção de atributos para o Bayes Net, Naive Bayes e SVM linear.

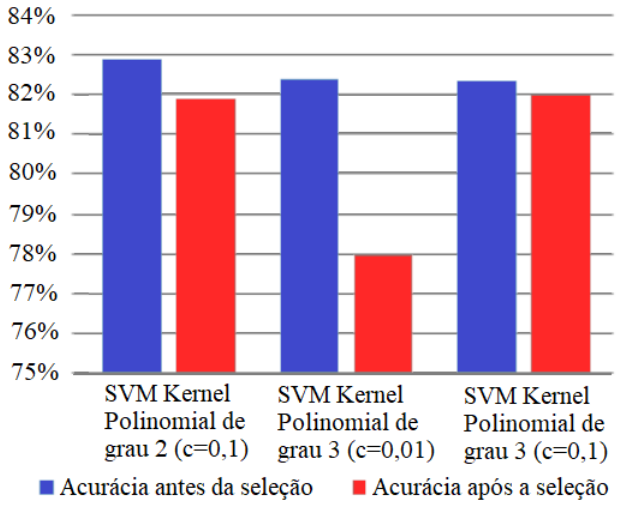

Fig. 2. Comparação entre acurácias antes e depois da seleção de atributos para os SVMs polinomiais de grau 2 e 3.

Ao utilizar a base integral obteve-se que os melhores resultados foram observados com o classificador *SVM Kernel Linear* com parâmetro de folga de 1,0. Por outro lado, após a redução de características, obteve-se que o classificador *Bayes Net* forneceu o melhor valor de acurácia.

Para o problema como um todo houve melhor classificação após seleção de atributos e por redes bayesianas. De fato, a redução de atributos (de 14 para 8, incluindo o atributo das classes) proporcionou condições mais favoráveis ao experimento, visto que a diminuição das características da base evita confusões do algoritmo de aprendizado, acelera o processo de aprendizado e permite que o classificador tenha foco nos atributos mais relevantes para o problema.

Além disso, como observado nas figuras, os valores das acurácias nas seis classificações elencadas como mais promissoras, em sua maioria, foram obtidos valores próximos (tanto utilizando a base integral com 14 atributos, assim como a base reduzida após a seleção de atributos). Nota-se, portanto, uma limitação em torno de 80% nos valores das acurácias. E, dado os devidos parâmetros de proporcionalidade, também foi verificado uma limitação nos valores de índice kappa, em torno de 0,6. O que já era esperado, pelo fato desses valores mostrarem-se condizentes entre si. É importante ressaltar os valores de sensibilidade e especificidade encontrados, correspondentes a bons resultados quanto a eficiência dos classificadores na predição, principalmente dos não doentes correspondentes a valores acima de 87% (especificidade). A área abaixo da curva ROC também se mostra com valores relevantes, no geral maiores que 0.80, apontando uma boa capacidade de classificar corretamente aqueles com e sem doença.

Acredita-se que a limitação dos parâmetros mencionados seja causada pela natureza da base utilizada. Uma das possibilidades é que as classes estejam discretamente desbalanceadas, o que gera uma tendenciosidade na classificação e, consequentemente, uma fronteira máxima nos valores que serão obtidos. Além disso, há a possibilidade da quantidade de dados presentes na base oferecerem apenas resultados margeando os números obtidos. Por fim, a natureza do problema pode ser a causa dessas limitações, o que pode ser verificado utilizando outras bases de dados.

IV. Conclusão

No projeto foram feitos experimentos para a predição de doenças cardíacas através da utilização de algoritmos classificadores e variando seus parâmetros, a fim de se observar a melhor predição. A partir dos resultados encontrados, verificou-se que o classificador Bayes Net aplicado na base de dados após a seleção de atributo obteve a maior acurácia de 83,143%. Ademais, é evidente a semelhança dos valores de acurácia para os classificadores, mesmo após a realização da seleção de atributos. Isso pode ocorrer devido a qualidade dos dados obtidos da base, fazendo-se obter uma acurácia limitada. Dessa forma, é viável a utilização de outras bases de dados, a fim de verificar a natureza da estagnação.

Portanto, uma vez que o objetivo do projeto é proporcionar uma forma de auxiliar o profissional de saúde na predição das doenças cardíacas, obteve-se resultados satisfatórios

Agradecimentos

A equipe dos docentes da disciplina de Inteligência Artificial Aplicada a Engenharia Biomédica do Departamento de engenharia Biomédica (DEBM) da Universidade Federal de Pernambuco (UFPE), por facilitarem e proporcionarem os momentos que permitiram a obtenção dos conhecimentos e incentivo na área de aprendizagem de máquina e inteligência artificial, além da disponibilização da base de dados utilizada neste trabalho.

Conflito de Interesse

Os autores declararam não ter havido conflitos de interesses.

Referências

[1]. Doenças Cardiovasculares. Organização Pan - Americana da Saúde, 2017. Disponivel em <https://www.paho.org/pt/topicos/doencas-cardiovasculares #:~:text=Dados%2FEstat%C3%ADsticas%3A,as%20mortes%20em%20n%C3%ADvel%20global>. Acesso em 07 de Nov. 2020.

[2]. Doenças Cardiovasculares (DCVs), World Health Organization (WHO). 2017. Disponível em <https://www.who.int/news-room/fact-sheets/detail/cardiov ascular-diseases-(cvds)>. Acesso 08 de Nov. 2020.

[3]. Prevenção e tratamento precoce são essenciais no combate às doenças do coração, Estúdio de SNC Total: 2020. Disponível em<https://www.nsctotal.com.br/noticias/prevencao-e-tratamento-precoce-sao-essenciais-no-combate-as-doencas-do-c oracao>. Acesso em 04 de Nov. 2020.

[4]. Machine Learning na saúde. Iclinic, 2018. Disponível em <https://blog.iclinic.com.br/machine-learning-na-saude-con heca-essa-revolucao/>. Acesso 03 de Nov. 2020.

[5].Saúde, Biologia, Classificação, Problemas de coração, Classificação binária, Kaggle, 2007. Disponível em <https://www.kaggle.com/ronitf/heart-disease-uci>. Acesso em 03 de Nov. 2020.

[6]. 1 Instituto Hungaro de Cardiologia. Budapeste: Janose Andras, MD. 2 University Hospital, Zurich, Suiça: Steinbrunn Willian, MD. 3 University Hospital, Basel, Suiça: Pfisterer Mathias, MD. 4 VA Medical Center, Long Beach e Cleveland Clinic Foundation: Detrano Robert, MD, Ph.D. Heart deseat data set. UCI Machine Learning Repository, 2007. Disponível em<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>. Acesso 03 de Nov. 2020.

[7]. Rawat, Shubhankar. Predição de doenças cardíacas. Towards data science, 2019. Disponível em <https://towardsdatascience.com/heart-disease-prediction-7 3468d630cfc>. Acesso em 04 de Nov. 2020.

[8] Mikhail, E. M., Ackermann, F. E. (1976). *Observations and least squares*. p. 64. New York, IEP.

[9] Coluna Psiquiatria Contemporânea. 2009. “Psiquiatria E Estatística V: Validação De Procedimentos Diagnóstica Pela Curva R.O.C”. Disponível em: <http://www.polbr.med.br/ano09/cpc0409.php>. Acesso em 07 de Set. 2021.

[10] Suporte ao Minitab® 18. 2019. “Estatísticas Kappa e coeficientes de Kendall”. Disponível em: <https://support.minitab.com/pt-br/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/> Acesso em 07 de Set. 2021.

[11] SILVA, G. S. L. E. ; OLIVEIRA, C. S. ; CAVALCANTI, L. H. ; BEZERRA, R. S. ; SILVA, I. R. R. ; SANTOS, W. P. . Sistema inteligente de apoio ao diagnóstico precoce da doença de Alzheimer usando análise multirresolução de imagens de ressonância magnética. In: II Simpósio de Inovação em Engenharia Biomédica - SABIO 2018, 2018, Recife. Anais do II Simpósio de Inovação em Engenharia Biomédica - SABIO 2018. Recife: BioTech Consultoria, 2018. p. 70-76.

[12] MAIA, M. ; PEQUENO, P. H. A. ; SILVA, W. W. A. ; SILVA, G. S. L. E. ; SANTANA, M. A. ; SANTOS, W. P. . Inteligência Artificial para o Apoio ao Diagnóstico da Doença de Alzheimer utilizando Imagens de Ressonância Magnética. In: Simpósio de Inovação em Engenharia Biomédica - SABIO 2019, 2019, Recife. Anais do Simpósio de Inovação em Engenharia Biomédica - SABIO 2019. Recife: BioTech Consultoria, 2019.

[13] OLIVEIRA, A. P. S. ; SANTANA, M. A. ; ANDRADE, M. K. S. ; GOMES, J. C. ; RODRIGUES, M. C. A. ; SANTOS, W. P. . Early Diagnosis Of Parkinson?S Disease Using Eeg, Machine Learning And Partial Directed Coherence. Research On Biomedical Engineering, V. 36, P. 311-331, 2020

[14] SOUZA, T. K. S. ; ANDRADE, J. F. S. ; ALMEIDA, M. B. J. ; SANTANA, M. A. ; SANTOS, W. P. . Métodos Computacionais Aplicados ao Diagnóstico de Câncer de Mama por Termografia: uma revisão sistemática. In: Simpósio de Inovação em Engenharia Biomédica - SABIO 2019, 2019, Recife. Anais do Simpósio de Inovação em Engenharia Biomédica - SABIO 2019. Recife: BioTech Consultoria, 2019.

[15] PEREIRA, J. M. S. ; SANTANA, M. A. ; LIMA, R. C. F. ; SANTOS, W. P. . LESION DETECTION IN BREAST THERMOGRAPHY USING MACHINE LEARNING ALGORITHMS WITHOUT PREVIOUS SEGMENTATION. In: Wellington Pinheiro dos Santos; Maíra Araújo de Santana; Washington Wagner Azevedo da Silva. (Org.). Understanding a Cancer Diagnosis. 1ed.New York: Nova Science, 2020, v. 1, p. 81-94.