CLASSIFICAÇÃO DE DOENÇAS CARDIOVASCULARES UTILIZANDO APRENDIZADO DE MÁQUINA

CLASSIFICATION OF CARDIOVASCULAR DISEASES USING MACHINE LEARNING

Marcos Costa Oliveira1
Luís Victor Belo Ferreira 2
Marta de Oliveira Barreiros 3

Resumo: A doença cardiovascular é a principal causa de morte para homens e mulheres, sendo conhecida por silenciosa. O diagnóstico sobre a predisposição para doenças cardiovasculares, mostra-se problemas que precisam de atenção e ajudam a alertar o paciente de possíveis complicações futuras. Nisso, o uso de algoritmos de aprendizado de máquina simplificara a intervenção humana em análises e tomadas de decisões, agilizando esses diagnósticos prévios para diferentes doenças. Portanto, este trabalho apresenta um comparativo entre três algoritmos de aprendizado de máquina, sendo eles o K-Nearest Neighbor (KNN), a Árvore de Decisões e a Rede Neural Multicamadas Perceptron (MLP), para escolha de um modelo com obtenção de melhores resultados para classificação de saúde em relação a doenças cardiovasculares. Diante do analisado, a Árvore de Decisão apresentou melhores resultados de classificação superior a 96% de precisão.

Palavras-chave: Classificação. Doenças Cardiovasculares. Aprendizado de Máquina.

Abstract: Cardiovascular disease is the leading cause of death for both men and women and is known to be silent. The diagnosis on the predisposition to cardiovascular diseases, shows up problems that need attention and helps to alert the patient of possible future complications. In this, the use of machine learning algorithms will simplify human intervention in analysis and decision-making, streamlining these previous diagnoses for different diseases. Therefore, this work presents a comparison between three machine learning algorithms, namely the K-Nearest Neighbor (KNN), the Decision Tree and the Multilayer Perceptron Neural Network (MLP), to choose a model with better results. for health classification in relation to cardiovascular diseases. In view of the analyzed, the Decision Tree presented better classification results than 96% of accuracy.

Keywords: Classification. Cardiovascular Diseases. Machine Learning.

1- Graduando em Engenharia da Computação pela Universidade Estadual do Maranhão. Email; marcostaoliv02@gmail.com. Lattes: http://lattes.cnpq.br/3294373593837147.

2 - Graduando em Engenharia da Computação pela Universidade Estadual do Maranhão. Email; luis.victor.belo@gmail.com. Lattes: http://lattes.cnpq.br/9572891935643593.

3 - Doutora em engenharia Elétrica pela Universidade Federal do Maranhão. Professora substituta no departamento de engenharia da Computação da Universidade Estadual do Maranhão. Lattes:http://lattes.cnpq.br/2695239794047991. ORCID: https://orcid.org/0000-0002-1367-1968. E-mail: barreiros1232@gmail.com.

Introdução

No Brasil, a doença cardiovascular (DCV) é a principal causa de morte e incapacidade para homens e mulheres. Entre as DCV, a doença isquêmica do coração (DIC) é a principal causa de morte no país, segundo estimativas do estudo Global Burden of Disease Study de 2019 (GBD, 2019). Além disso, as DCV também são conhecidas por serem silenciosas causando óbitos fatais. Diante disso, diagnósticos sobre a predisposição para doenças cardiovasculares, mostram-se problemas que precisam de atenção e ajudam a alertar o paciente de possíveis complicações futuras.

A cada evolução humana, é inegável o crescimento exponencial de dados que nos cercam e transformam nossa maneira de viver. Diante disso, o crescimento de técnicas e ferramentas para análise e manipulação desses dados fazem-se necessárias. A Inteligência Artificial (IA), vem ganhando forças e se adequando a diferentes problemas cotidianos, visto que é uma área da tecnologia que consiste em técnicas computacionais que se baseiam no comportamento humano para resolver problemas. Ela tem o intuito de tornar capaz um computador a analisar dados, encontrar padrões ou tendências e assim tomar uma decisão. Essas tomadas de decisões tendem a melhorar quanto mais o computador é treinado para resolver um problema determinado. Estes treinamentos são realizados utilizando algoritmos específicos para cada propósito: classificação, predição, agrupamento, etc.

Ademais, *Machine Learning* (ML), Aprendizado de Máquina em tradução livre, é um ramo da inteligência artificial que consiste em modificar o comportamento autonomamente do computador tendo como base uma experiência "vivida". Dessa forma, apresenta-se como uma solução promissora, para simplificar a intervenção humana em análises e tomadas de decisões. Entretanto, determinar o ML que melhor se adéque ao problema dentre as possibilidades, requer atenção e cautela.

Neste trabalho, aplicou-se os algoritmos de *Machine Learning* KNN e Árvore de Decisão e Rede Neural Multicamadas Perceptron (MLP), em uma base de dados com informações biológicas de pacientes para classificar aqueles com problemas cardíacos auxiliando futuros diagnósticos para a predisposição para doenças cardiovasculares.

Materiais e metodologia

Este trabalho implementou os três algoritmos, na linguagem *Python* e utilizou a ferramenta Google Collab, como Ambiente Integrado de Desenvolvimento (IDE), por ser uma plataforma online que permite um trabalho em grupo mais ágil e também é mais simples na importação e na instalação de bibliotecas. Também foram usadas as bibliotecas sklearn (PEDREGOSA *et al.,* 2011) e pandas (MCKINNEY, 2010), para implementação dos algoritmos de classificação e para manipulação e análise de dados, respectivamente.

Base de dados

O estudo utilizou a base de dados pública heart.csv (DAVID W., 2019), sobre doenças cardiovasculares. Para alcançar os objetivos propostos foi modelada a classificação de saúde em relação a doenças cardiovasculares. A base conta com um conjunto de dados de mil e vinte cinco pacientes (1025), nos quais 526 apresentam doença cardiovascular e 499 não possuem doença cardiovascular. Além disso, o arquivo .csv apresenta dados de 13 características, sendo elas idade, sexo, tipo de dor no peito (cp), pressão arterial em repouso (trestbps), colesterol, glicemia, resultados de eletrocardiograma em repouso, frequência cardíaca máxima, angina induzida por exercício, depressão de ST induzida por exercício em relação ao repouso, Inclinação do segmento ST de pico do exercício, número de vasos principais coloridos por fluoroscopia e talassemia. Foi utilizada a base completa com todas as colunas de dados, por preferência do

autor, para que a comparação seja feita com a base bruta para que seja observado a precisão dos algoritmos com dados considerados ruins para o resultado final.

Pré-processamento dos dados

Devido a existência de variação na forma que esses dados são expressos no arquivo, os mesmos passaram por uma normalização antes da implementação dos algoritmos de classificação. Em seguida, para evitar *overfitting* e *underfitting* no modelo foi aplicado a técnica de validação cruzada K-Fold, na qual subdivide o conjunto de dados em subconjuntos, de mesmo tamanho, assim, um subconjunto é utilizado de teste e os demais para estimação do modelo. O processo é realizado K vezes, sendo k o número de *folds* definidos. Logo, o seu custo computacional varia de acordo com o tamanho do *dataset* selecionado e no fim retorna um score da performance de cada subdivisão. Esse resultado é aplicado nos três algoritmos analisados, com a utilização de 3 *folds*.

Medidas de avaliação

Posteriormente, foi obtido uma saída com parâmetros métricos (PEDREGOSA *et al.,* 2011), nos quais são medidas estatísticas comumente utilizadas pela comunidade, sendo eles *precision*, *recall*, *f1-score* e *support*.

$$Precision: \quad \frac{tp}{tp + fp} \qquad (1)$$

$$Recal: \quad \frac{tp}{tp + fn} \qquad (2)$$

$$F1 - score: \quad \frac{2 \cdot precision \cdot recall}{precision + recall} \qquad (3)$$

em que tp é o número de verdadeiros positivos, fp o número de falsos positivos e fn o número de falsos negativos. Dessa maneira, *precision* é a capacidade do classificador de analisar positivos corretamente, ou seja, seja sem rotular uma amostra negativa como positivo. Além disso, o *recall* indica o quanto nosso modelo está identificando os casos positivos corretamente, encontrando todas as amostras positivas. A *F1-score* é uma maneira de visualizarmos as métricas *precision* e *recall* juntas, por meio de uma média harmônica ponderada. Por fim, nos resultados, também é apresentada a métrica *support*, na qual indica o número de dados ocorrentes.

K-Nearest Neighbor (KNN)

O KNN (*K-Nearest Neighbours*) é um algoritmo que classifica um dado com base na maior quantidade de vezes que uma classificação se faz presente entre K vizinhos mais próximos a este mesmo dado. Os K vizinhos mais próximos são aqueles que têm menor distância (podendo ser euclidiana, ponderada, etc.) do dado classificado. O estudo utiliza 3 vizinhos (K) na implementação.

Árvore de decisão

A Árvore de Decisão é um algoritmo que classifica um dado criando uma estrutura de árvore que cada nó interno é um teste em um atributo, cada ramo representa um resultado do teste e cada nó folha armazena um rótulo de classe, ou classificação final. Para criar esta estrutura, ela realiza o cálculo de entropia e de ganho para definir o melhor atributo, o qual

será o nó raiz, seguido dos nós internos até chegar nos nós folhas. À árvore de decisão foram testadas apenas as configurações de variáveis de entrada no algoritmo.

Rede Neural Multicamadas Perceptron (MLP)

Nas Redes MLP, há a implementação de uma camada oculta e a utilização de backpropagation para o treinamento da rede em sua arquitetura. Nessas redes, cada camada tem uma função específica. A camada de saída recebe estímulos das camadas intermediárias/ocultas e constrói os padrões que se tornarão as respostas. As camadas intermediárias servem como extratores de recursos, e seus pesos são uma codificação dos recursos apresentados no padrão de entrada e permitem que a rede crie sua própria representação mais rica e complexa do problema (HAYKIN, 2007).

Dessa maneira, o estudo utiliza das configurações de 2 camadas intermediárias de 600 e 200 neurônios, respectivamente, e com função sigmoide para a ativação dos neurônios. Além disso, para o aprendizado, teve a taxa de aprendizado de 0,00001, 500 epochs e otimizador adam.

Resultados e discussões

Nesta seção, é apresentado um resumo sobre os resultados alcançados pela análise comparativa entre os três algoritmos, após a modelagem dos dados, como exposto na seção anterior e sendo, 0 é pessoa saudável e 1 pessoa que apresenta doença cardiovascular. Dessa forma, é apresentado na Tabela 1(KNN), Tabela 2 (Árvore de Decisão) e Tabela 3 (Rede MLP).

Tabela 1. Resultados KNN com aplicação de 3 *folds*

Fold	Classe	Precision	Recall	F1-score	Support
01	0	0,90	0,92	0,91	158
	1	0,93	0,91	0,92	184
02	0	0,92	0,81	0,86	170
	1	0,83	0,93	0,88	172
03	0	0,86	0,92	0,89	171
	1	0,91	0,85	0,88	170

Fonte: Marcos Costa Oliveira (2022).

Tabela 2. Resultados Árvore de Decisão com aplicação de 3 *folds*

Fold	Classe	Precision	Recall	F1-score	Support
01	0	0,96	1	0,98	158
	1	1,00	0,97	0,98	184
02	0	1,00	0,94	0,97	170
	1	0,95	1,00	0,97	172
03	0	0,97	0,98	0,97	171
	1	0,98	0,96	0,97	170

Fonte: Marcos Costa Oliveira (2022).

Tabela 3. Resultados Rede MLP com aplicação de 3 *folds*

Fold	Classe	Precision	Recall	F1-score	Support
01	0	0,88	0,81	0,84	158
	1	0,85	0,91	0,88	184
02	0	0,88	0,73	0,8	170
	1	0,77	0,90	0,83	172
03	0	0,85	0,68	0,76	171
	1	0,74	0,88	0,80	170

Fonte: Marcos Costa Oliveira (2022).

Vale pontuar que, a Árvore de Decisão (Tabela 2), estava tendendo ao *overfitting* em alguns momentos, mas o mesmo foi reduzido nos outros *folds*, por conta da utilização da validação cruzada no pré-processamento dos dados.

O resultado final é a média dos valores encontrados após cada validação/*fold*, como representado na Tabela 4, que contém os três métodos. Assim, analisando os modelos de forma individual, é possível perceber que dentre os modelos, a Árvore de Decisão apresentou-se mais eficiente, visto que está indicando com maior certeza e precisão os casos positivos para doenças cardiovasculares, logo promovendo um melhor diagnóstico.

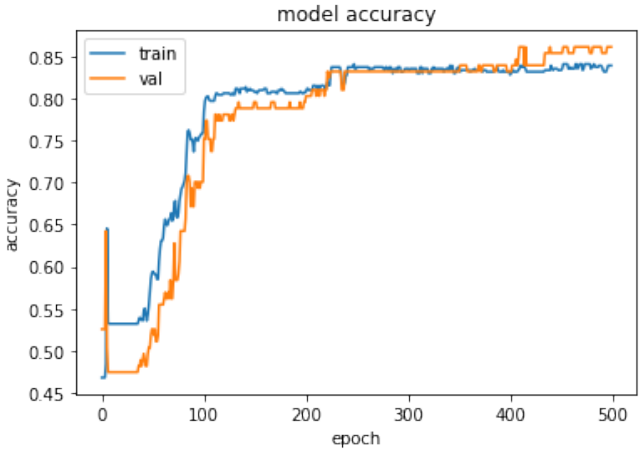
Tabela 4. Média dos Resultados KNN, Árvore de Decisão e Rede MPL com aplicação de 3 *folds*

	Classe	Precision	Recall	F1-score	Support
KNN	0	0,90	0,92	0,89	170,00
	1	0,91	0,91	0,88	172,00
Árvore de Decisão	0	0,97	0,98	0,97	170,00
	1	0,98	0,97	0,97	172,00
Rede MLP	0	0,88	0,73	0,80	170,00
	1	0,77	0,90	0,83	172,00

Fonte: Marcos Costa Oliveira (2022).

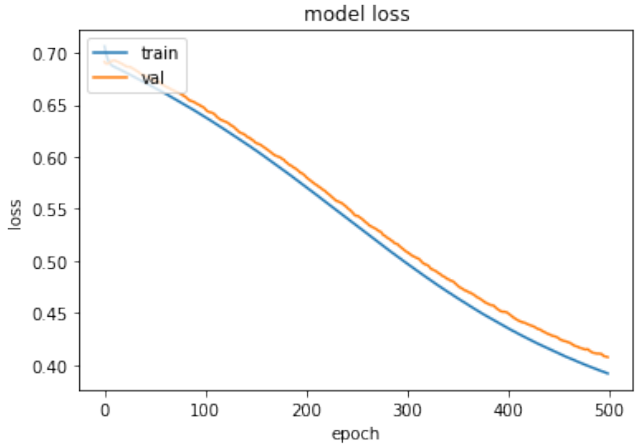
Enquanto a Rede MLP, mesmo mostrando bons resultados, com valores consideráveis para F1-score dentre os modelos comparados apresentou resultados mais baixos. O fato pode ser observado devido a utilização de uma base de dados pequena e do *backpropagation* que o modelo possui, para aplicação de MLP, que normalmente é utilizada em problemas com grande custo computacional. As Figuras 1 e 2, representam o supracitado em relação a outras métricas, sendo a acurácia que representa de um modo generalizado a *performance* do modelo e a perda, que representa quanto é a diferença entre o valor de saída da rede e o esperado para uma determinada entrada.

Figura 1. Acurácia do Modelo X Epoch



Fonte: Marcos Costa Oliveira (2022).

Figura 2. Perda do Modelo X Epoch



Fonte: Marcos Costa Oliveira (2022).

Dessa maneira, percebe-se que a rede não apresentava uma perca tão gradativa, para melhorar a acurácia do modelo. Além disso, o KNN demonstrou resultados de aprendizado próximos ao da Árvore de Decisão, o que confirma a utilização do método em problemas simplificados.

Considerações Finais

O presente trabalho objetiva estudar três algoritmos de *machine learning,* para classificação de saúde em relação a doenças cardiovasculares, a fim de compreender os métodos e comparar o desempenho para o conjunto de dados utilizado.

Com os resultados obtidos, pode-se perceber que tanto o modelo de Árvore de Decisão quanto o KNN podem ser utilizados para a realização de classificação para doenças cardiovasculares. Com esses resultados, o estudo objetiva auxiliar futuros diagnósticos para a predisposição para doenças cardiovasculares. Ademais, o estudo permite novos direcionamentos para trabalhos futuros, como a ampliação de dados medicinais referentes a diagnósticos de doenças cardiovasculares, para busca de melhor modelo que possua um maior desempenho para a classificação.

Ainda assim, dia após dia essa área vem ganhando espaço no nosso cotidiano, deixando

de ser um assunto abordado em obras de ficção científica e passando a ser presente na nossa própria casa. Logo, é importante que haja cada vez mais pesquisas e estudos a cerca deste ramo da tecnologia, afim de tornar a vida do ser humano mais prática e de maior qualidade.

Referências

DAVID W. **Heart Disease Data set exploration, Kanggle**, 2019. Disponível em: https://www.kaggle.com/datasets/volodymyrgavrysh/heart-disease. Acessado dia 01 nov. 2022.

GBD. Global Burden of Disease Study 2019. **Global Health Data Exchange website** [Internet]. Seattle, WA: Institute for Health Metrics and Evaluation (IHME), University of Washington; 2019. Acesso em: http:// ghdx.healthdata.org/gbd-results-tool

HAYKIN, S. **Redes Neurais**: Princípios e Prática. Artmed, 2007. ISBN 9788577800865isp.

MCKINNEY, **Data structures for statistical computing in python**, Proceedings of the 9th Python in Science Conference, Volume 445, 2010.

PEDREGOSA *et al.*, Scikit-learn: **Machine Learning in Pytho**n, JMLR 12, pp. 2825-2830, 2011.

PEDREGOSA *et al.,* **Developer, scikit-learn.** Sklearn.metrics.precision_recall_fscore_support. Disponível em: sklearn.metrics.precision_recall_fscore_support — documentação scikit-learn 1.1.2. Acessado dia 01 nov. 2022.

Recebido em: 30 de novembro de 2022.
Aceito em: 20 de janeiro de 2023.