Um Método Baseado em Aprendizado de Máquina para Detecção de Patologias Cardíacas em Pacientes Pediátricos

Tiago Carlos Batista
Departamento de Computação (DC)
Universidade Federal de São Carlos (UFSCar)
13566-447, São Carlos, Brasil
tiagodr07@gmail.com

***Resumo*—As doenças cardíacas pediátricas, tanto congênitas quanto adquiridas, são causas importantes de morbidade e mortalidade. O diagnóstico precoce é essencial, mas desafiador em ambientes com recursos limitados. Este estudo propõe o uso de Aprendizado de Máquina (AM) para a detecção de patologias cardíacas em crianças, utilizando uma base de dados do Real Hospital Português (RHP). Foram aplicados os algoritmos *k*-vizinhos próximos (KNN), *Naive Bayes* (NB), Regressão Logística (RL), Redes Neurais Artificiais (RNA) e Máquinas de Vetores de Suporte (do inglês, *Support Vector Machine* - SVM). A avaliação dos modelos foi feita com as métricas AUC, Acurácia, Precisão, Recall e *F1-Score*. O modelo de RNA obteve a melhor performance, com uma AUC de 0.9505, mostrando a eficácia do AM no diagnóstico de doenças cardíacas pediátricas.**

***Palavras-chave*-Doenças cardíacas; diagnóstico precoce; aprendizado de máquina; redes neurais artificiais; classificação, AUC; pediatria**

I. Introdução

As doenças cardíacas congênitas e adquiridas são causas relevantes de morbidade e mortalidade na população pediátrica [1]. O diagnóstico precoce dessas condições é crucial para garantir tratamento adequado e melhorar a qualidade de vida dos pacientes, mas pode ser desafiador, especialmente em contextos com recursos limitados e poucos especialistas.

Avanços em Aprendizado de Máquina (AM) oferecem novas possibilidades para o diagnóstico médico. Modelos de AM, treinados para identificar padrões em exames clínicos como eletrocardiogramas (ECG) e ecocardiogramas, contribuem para avaliações mais rápidas e precisas. Modelos de *Deep Learning* já demonstraram alto desempenho na detecção de estenose aórtica e regurgitação mitral [2]. Neste contexto, este trabalho propõe o uso de modelos de AM para a detecção de patologias cardíacas em pacientes pediátricos.

A pesquisa utiliza dados do Real Hospital Português (RHP), em parceria com a Universidade do Porto, visando avaliar a eficiência de algoritmos de AM na classificação de pacientes saudáveis e com patologias cardíacas. Foram aplicados os algoritmos *k*-vizinhos próximos (KNN), *Naive Bayes* (NB), Regressão Logística (RL), Redes Neurais Artificiais (RNA) e Máquina de Vetores de Suporte (SVM), implementados com a biblioteca *Scikit-learn*. As métricas de avaliação incluem AUC (Área Sob a Curva ROC), Acurácia, Precisão, Recall e *F1-Score*, com destaque para a RNA, que obteve AUC de 0.9505.

Este artigo está organizado da seguinte forma: a Seção II discute trabalhos relacionados; a Seção III descreve as bases de dados e o pré-processamento; a Seção IV detalha o protocolo experimental, incluindo avaliações e ajustes de parâmetros; a Seção V apresenta e analisa os resultados, incluindo o desempenho no *Public Leaderboard* do *Kaggle*; a Seção VI aborda as estratégias finais da competição; e a Seção VII conclui com um resumo dos resultados e sugestões para trabalhos futuros.

II. Trabalhos Relacionados

Nos últimos anos, pesquisas têm explorado o uso de AM no diagnóstico de doenças cardíacas, mostrando que algoritmos avançados aumentam a precisão na detecção de anomalias e reduzem a dependência de métodos manuais. Esta seção discute trabalhos da literatura sobre técnicas de AM para classificar exames cardíacos, destacando suas principais contribuições.

Uma pesquisa sobre farmacoterapia em pacientes pediátricos transplantados cardíacos analisou dados de 18 pacientes (1 a 18 anos), identificando alterações em transaminases (27,8%), ureia e creatinina (33,3%) e anemia (50%) pós-transplante, com uso universal de estatinas como profilático [3]. O estudo destaca a importância do acompanhamento contínuo e intervenções para melhorar os resultados. O transplante, embora vital para insuficiência cardíaca terminal, exige cuidados médicos regulares devido à escassez de doadores compatíveis, especialmente em neonatos. Nesse contexto, métodos de AM podem auxiliar no diagnóstico precoce, contribuindo para o sucesso do tratamento.

Um estudo com o objetivo de construir um modelo de AM profundo para classificar sons cardíacos como normais ou anormais usou o conjundo de dados do concurso *George B. Moody PhysioNet Challenge*. O trabalho aplicou diferentes segmentações de áudio, extraiu características de espectrogramas de mel e MFCC e utilizou algoritmos de

aprendizado profundo, alcançando uma acurácia de 77,36% e um *F1-Score* de 62,22% [4]. O mesmo estudo ressalta que as doenças cardiosvasculares foram responsáveis por 33% das mortes em 2019, e não menos importante, a dificuldade em diagnosticar essas doenças devido à ausência de sintomas evidentes [4]. Soluções que auxiliem profissionais de saúde a realizar o rastreio inicial de anomalias cardíacas, também são preteridas neste trabalho, assim como no anterior, ratificando como o AM pode ser útil nesse aspecto.

A Síndrome Metabólica (SM) sucedida em crianças obesas apresentam fatores de risco para doenças cardiovasculares, impulsionadas pela resistência à insulina causada pela SM [5]. Os principais métodos utilizados para avaliar a obesidade e a resistência à insulina nas crianças do estudo foram:

- **Avaliação do Índice de Massa Corporal (ICM):** A obesidade foi definida quando o IMC estava acima do percentil 95 de acordo com a classificação do *Centers for Disease Control and Prevention* (CDC). O IMC foi calculado mediante a Equação (1):

$$IMC = \frac{\text{massa (kg)}}{\text{altura (m)}^2} \tag{1}$$

- **Medições Antropométricas:** Foram realizadas medições de peso e altura com equipamento apropriados, e também foi obtida a circunferência da cintura na altuira da cicatriz umbilical;
- **Exames de Sangue:** O estudo utilizou amostras de sangue de jejum para mensurar a glicemia, níveis de lipoproteínas de alta densidade (HDL), triglicerídeos e insulina;
- **Avaliação da Resistência à Insulina:** A resistência à insulina foi determinada pelo método de homeostase glicêmica (HOMA-IR), que é o produto da insula de jejum e da glicemia de jejum divido por 22,5.

Detecções precoces de doenças cardiovasculares em crianças também pode ser um diferencial neste contexto apresentado.

III. Dados e Pré-processamento

Os dados oriundos do RHP, a priori, foram divididos em três partições: "RHP_data.csv", "train.csv" e "test.csv". Foram respectivamente renomeados para "Base_Principal.csv", "Base_Treino.csv" e "Base_Teste.csv". A base de dados intitulada como "Base_Principal" é onde estão todos os dados referentes aos pacientes, uma espécie de prontuário médico que carrega informações como peso, altura, IMC, entre outros. As outras duas bases, respectivamente, se tratam de dados para uso no momento de treino e teste dos modelos propostos.

Em um primeiro momento, na análise exploratória, as operações se limitaram em verificações de registros nulos e negativos para colunas ou *features* que já carregam seus registros em formato numérico, e em casos de *features* categóricas, além das verificações anteriores, também foi consultado o tipo de registro existente e/ou seus valores únicos. Por exemplo, nas *features* "Peso" e "Altura", as únicas técnicas de análises foram verificações que retornaram possíveis valores nulos ou negativos, já na *feature* "Atendimento" (Data de Atendimento) e "DN" (Data de Nascimento), uma verificação mais específica se mostrou necessário, já que os registros supostamente possuem um tipo diferente do numérico. Somente duas execções foram detectadas onde se foi verificado o tipo de dado em *features* com valores aparentemente numéricos, na *feature* "IDADE", onde foi verificado o tipo de registro da *feature*, retornando *"object"* e também na *feature* "FC" (Frequência Cardíaca). Nesse momento da análise exploratória, "IDADE" já sofreu um processamento, onde todos os seus dados foram convertidos para o tipo numérico e forçando possíveis erros em valores a se tornarem valores nulos.

Iniciando o pré-processamento, a primeira feature a ser manuseada foi a "Peso". Na análise exploratória, foi constado registros com valores negativos e os mesmos foram substituídos pelo valor nulo. Todos os valores nulos da coluna foram substuídos pela mediana dos valores da coluna. Por fim, também foi detectado registros com valor 0 na coluna, essa informação será importante para operações de pré-processamento em um momento posterior, com outra *feature*. Em relação a coluna "Altura", na análise exploratória não foram encontrados registros que possuíssem valores nulos ou negativos, portanto em seu pré-processamento também foram detectados registros zerados, assim como na coluna anterior, e que será importante para a coluna a seguir. Na coluna "IMC", não foram encontrados valores negativos, somente nulos. Tais valores nulos foram substuídos através da equação (1), onde a massa foi substuído pelo valor correspondente do registro da coluna "Peso" e altura pelo valor correspondente do registro da coluna "Altura", respectivamente. Depois desse tratamento, a quantidade de valores nulos da coluna "IMC" diminuiu de 4.727 registros para 2.500 registros, indicando ainda que precisaria de algum tratamento. Por intuição, desconfiou-se que esses 2.500 registros sendo acusados, seriam resultado de divisões por 0 oriundos do tratamento utilizando a equação (1) ou quando os valores das duas colunas "Peso" e "Altura" fossem 0. Portanto, verificou-se a quantidade de registros onde "Peso" é 0, "Altura" é 0 e "IMC" é nulo, e essa quantidade deu exatamente 2.500 registros. Concluiu-se então que, esses valores, deveriam ser alterados para 0, seu resultado legítimo. Seguindo para as colunas "Atendimento" e "DN", de primeiro instante, foi tentado uma forma manual de encontrar uma possível relação entre as duas, juntamente com a coluna "IDADE", como por exemplo, realizar uma "diferença" entre registros de "Atendimento" e "DN" e encontrando um dado que faça sentido para a coluna "IDADE". Porém, não foi encontrado. Sendo assim, uma solução prevista foi realizar a ténica de correlação entre "Atendimento"

e ”DN” e a partir dos resultados, foi apurado uma correlação positiva fraca entre as duas *features*, desse modo, as duas colunas foram excluídas. Se tratando da coluna ”IDADE”, registros nulos detectados foram alterados para o valor da mediana da coluna, assim como registros negativos, porém neste caso, os valores previamente negativos não foram considerados para o cálculo da mediana. A *feature* ”Convenio” também sofreu uma análise mais ”intuitiva”, assim como feito com *features* anteriores, mas semelhante a resultados já alcançados, informações o tipo de convênio dos pacientes não seria útil nesse contexto do estudo e por isso, seria excluído. Na primeira *feature* categórica no pré-processamento, a ”PULSOS”, se tem uma operação que será muito frequente nesse tipo de coluna: o mapeamento das categorias existentes para o tipo numérico. Todas as categorias recuperadas, em primeiro momento, sofreram uma operação de *lower case* para facilitar a conversão e lidar com categorias com diferentes tipos de apresentações. Foram reconhecidas as categorias ”normais”, ”amplos”, ”femorais diminuídos” e ”outro”, todas convertidas para a numeração de 0 até 3, respectivamente. E por fim, os valores nulos presentes nessa coluna, foram substituídos pela moda calculada da mesma, que neste caso resultou na categoria 0. Seguindo para as features ”PA SISTOLICA” e ”PA DIASTOLICA”, pressão arterial sistólica e pressão arterial diastólica, respectivamente, os valores nulos foram substuídos através do cálculo da mediana de cada coluna. ”PPA” é uma outra *feature* que faz parte da base de dados, que se refere a pressão de pulso alongada. Nessa coluna, existem registros preenchidos como ”#VALUE!”, algum tipo de erro de incompreensão de dado, esses valores foram substuídos para registros nulos, As categorias reconhecidas, ”Não calculado”, ”Normal”, ”Pre-Hipertensão PAS”, ”HAS-2 PAS”, ”Pre-Hipertensão PAD”, ”HAs-1 PAS”, ”HAS-2 PAD” e ”HAS-1 PAD” foram mapeadas para as numerações de 0 até 7, respectivamente. Os valores nulos foram alterados para o cálculo da moda de coluna, ou seja, categoria 0. A próxima coluna para o pré-processamento, é a ”B2”, que supostamente está relacionado à caracterização de sopros ou sons cardíacos anormais. Nesta *feature*, se tratando de registros categóricos, foram reconhecidas como ”Normais”, ”Desdob fixo”, ”Outro”, Hiperfonética” e ”Única”, mapeados para a numeração de 0 até 4, respectivamente. Assim como nos pré-processamentos similares, valores nulos foram substuídos pelo cálculo da moda da coluna, que resultou na categoria 0. A próxima coluna, ”SOPRO”, o mapeamento feito continha dados redundantes, com mesmas categorias apresentados de jeitos diferentes. As categorias reconhecidas, ”ausente”, ”Sistólico”, ”sistólico”, ”Diastólico”, ”Contínuo”, ”contínuo” e ”Sistolico e diastólico” ganharam novos rótulos numéricos, de 0 até 4, pelo fato dos termos ”Sistólico” e ”Contínuo” apresentarem mais de uma maneira de serem escritas. Os valores nulos foram alterados para o cálculo da moda, que assim como nas *features* anteriores, resultou na categoria 0. Em relação a coluna ”FC” (Frequência Cardíaca), anteriormente já discutida, notou-se que o tipo de registros contídos na coluna não era numérico, e a primeira operação feita foi a alteração para numérico. Ao analisar os tipos únicos presentes na coluna, tanto valores muito baixos, quanto muito altos foram detectados e lidados como *outliers*. Esses valores que fogem da margem comum foram alterados para valores nulos, aumentando a quantidade de valores nulos presentes ness coluna de 2.041 para 2.075 registros. Como relatado, os valores únicos foram consultados, e além da análise de outliers feita, também foi detectado outro comportamento que deveria passar por um tratamento: valores com intervalos. Exemplo, a coluna ”FC” continha valores como ”50-100”, comum em medições caríacas, mas que pode gerar algum tipo de ruído para a o processamento da base de dados. Para lidar com esse tipo de dado, criou-se uma função que captura esses valores com intervalo, realiza a média entre essa faixa de valores e a substitui pelo resultado. Por exemplo, o dado ”50-100” foi alterado para ”75”. E os registros nulos, foram alterados para a mediana da coluna. Seguindo para a próxima coluna, se tem a ”HDA 1”, supostamente se trata da História da Doença Atual, que é um registro detalhado e cronológico dos síntomas que levaram um paciente a procurar a assistência médica. As categorias resgatadas dessa *feature* foram ”Palapitação”, ”Dispneia”, ”Assintomático”, ”Dor precordial”, ”Desmaio/tontura”, ”Outro”, ”Cianose” e ”Ganho de peso”, todas mapeadas para numeração de 0 à 7, respectivamente. Valores nulos foram alterados para o resultado do cálculo da moda da coluna: a categoria 0. A sua coluna ”irmã”, a ”HDA2”, apresentou número elevado de registros nulos, mais especificamente 17.221 registros nulos foram resgatados, sendo assim, para evitar problemas maiores de outliers, a *feature* ”HDA2” foi removida. A *feature* ”SEXO”, também categórica, foi mapeada. O sexo masculino foi representado de três maneiras distintas: ”M”, ”Masculino” e ”masculino”. Já a categoria de sexo feminino, duas maneiras distintas: ”F” e Feminino”. Além disso, outra categoria foi reconhecida, a ”Indeterminado”, sendo assim, o mapeamento foi conduzido de maneira que o sexo feminino fosse representar o valor 0, o sexo masculino valor 1 e a categoria ”Indeterminado”, 2. Como de costume nesse trabalho, os valores nulos foram substuídos pelo cálculo da moda da coluna. As duas últimas colunas da base, ”MOTIVO1” e ”MOTIVO2” seus registros já possuiam um valor numérico atrelado, por exemplo, a coluna ”MOTIVO1” apresentou as seguintes categorias: ”6 - Suspeita de cardiopatia”, ”2 - Check-up”, ”5 - Parecer cardiológico”, ”1 - Cardiopatia já estabelecida” e ”7 - outro”. Levando isso em conta, o mapeamento seguiu exatamente essa numeração já estabelecida, com os valores nulos substuídos pelo cálculo da moda, que nesse caso resultou na categoria 5. Encerrando o pré-processamento, a *feature* ”MOTIVO2” possui as categorias ”6 - Palpitação/taquicardia/arritmia”, ”6 - Dispnéia”,

”5 - Atividade física”, ”5 - Cirurgia”, ”6 - Sopro”, ”1 - Cardiopatia adquirida”, ”1 - Cardiopatia congenica”, ”6 - Dor precordial”, ”6 - HAS/dislipidemia/obesidade”, ”6 - Cianose”, ”Outro”, ”6 - Alerações de pulso/perfusão”, ”6 - Cansaço”, ”5 - Uso de cisaprida” e ”6 - Cianose e dispnéia”. Foi observado essa coluna é como se fosse uma extensão da anterior, apresentando as mesmas numerações vistas. Nessa coluna, o mapeamento feito na coluna anterior foi replicado, onde a numeração acompanhada pela nomeação, foi mantida. Valores nulos foi alterados para o cálculo da moda da coluna, a categoria 5. A base de treino, ”Base_Treino” também passou por pré-processamento. Ela conta com duas colunas, ”Id” e ”CLASSE”. A coluna ”CLASSE” continha categorias ”Normal”, ”Anormal” e ”Normais”, além de valores nulos. As categorias ”Normal” e ”Normais” foram somadas e os valores nulos modificados para a moda da coluna: a classe ”Normal”.

IV. Protocolo Experimental

Nesta seção, serão apresentadas as formas de avaliação, bem como ajuste de parâmetros dos modelos e das medidas de desempenho empregadas.

Todos os modelos descritos que foram usados para desempenhar, passaram pela mesma fase de preparação dos dados, onde foi feita a combinação da ”Base_Principal” com ”Base_Treino”, usando o ”Id” como chave. Para normalização dos dados, as métodos *StandardScaler* e *RobustScaler* foram usados. Para a validação cruzada, a *StratifiedKFold* auxiliou, variando os *folds* de 5 até 10. A principal medida de desempenho analisada, foi a AUC. Em relação aos algoritmos, o KNN por exemplo, os parâmetros testados são k de valendo 5, 9, 11, 15 e 21. Já na RNA, os números de camadas ocultas foram se alterando, assim como número de neurônios em cada uma delas, função de ativação ”ReLU” e ”Tanh” também foram testadas. Por fim, o modelo prevê a probabilidade de ser da classe ”Anormal” e salva os resultados em um CSV com intitulação personalizada para cada modelo.

V. Resultados

Nesta seção, serão apresentados os resultados obtidos das execuções dos modelos, além de variações do mesmo para alcançar pontuções no Public *LeaderBoard* da competição. Por isso, a Tabela 1 a seguir exibe os resultados obtidos nas execuções no *Google Colab*.

Tabela I
Resultados das Medidas de Desempenho dos Modelos

Modelo	Medidas de Desempenho				
	AUC	Acurácia	Precisão	Recall	F1-Score
KNN	0.9453	0.9300	0.9391	0.9200	0.9300
NB	0.9365	0.9211	0.9033	0.8799	0.8914
RL	0.9460	0.9357	0.9527	0.8685	0.9086
RNA	0.9505	0.9319	0.9391	0.8723	0.9085
SVM	0.9414	0.8962	0.9547	0.7539	0.8425

Os resultados da Tabela 1 contém as seguintes configurações de parâmetros: KNN - número de *k*-vizinhos = 21, com *StandardScaler* e *folds* = 5; NB - *RobustSacler* e *folds* = 5; RL - *penalty* = ’l2’, *solver* = ’liblinear’, com *RobustSacler* e *folds* = 5; RNA - *hidden_layer_sizes* = (512, 256, 128), *activation* = ’tahn’, *solver* = ’adam’, *learning_rate_init* = 0.001, *alpha* = 0.0005, max_iter = 500, com *RobustSacler* e *folds* = 10; SVM - *kernel* = ’rbf’, C = 1.0, com *RobustSacler* e *folds* = 5.

Já competição do *Kaggle*, no *LeaderBoard*, o primeiro modelo submetido foi com o KNN, com *k*-vizinhos = 5, obtendo resultado de 0.92709, lembrando que na competição, a métrica de desempenho adotada foi a AUC. As seguintes submissões foi com o SVM e RNA, com *kernel* ”RBF” e com 2 camadas ocultas (100, 50), com função de ativação ”ReLU” e otimizador ”Adam”, com pontuações 0.92533 e 0.93292, respectivamente. Depois disso, variações de parâmetros foram testados com a RNA. Logo em seguida, a RL demonstrou um desempenho de 0.93394, e o NB 0.93411. Vale ressaltar que as diferenças entre as pontuções de AUC da Tabela 1 e dos resultados das submissões para a competição, são causadas pelo fato dos conjuntos de teste serem distintos entre as execuções do *Colab* e do *Kaggle*.

VI. Estratégia Final

Para a estratégia final, no que diz respeito a competição, foi percebido que a RNA se mostrou muito eficiente, pois foi com esse modelo que se ultrapassou as pontuções de todos os três *benchmarks*. Se trata da terceira RNA com parâmetros alterados, que resultou na primeira grande pontuação de 0.94215 na competição, que contou com 3 camadas ocultas (128, 64, 32) ”ReLU”, ”Adam” e 500 iterações.

Alguns algoritmos bônus foram testados, a *Random Forest* e o *Gradient Boosting Classier*, porém ambos não desempenharam melhor que a RNA na competição, com pontuções de 0.93938 e 0.50202, respectivamente. A segunda grande pontuação alcançada, também foi com a RNA, com pontuação de 0.94341, e os parâmetros: 3 camadas ocultas (256, 128, 64) função de ativação Tahn e 500 iterações.

VII. Conclusão

Os resultados obtidos, de modo geral, se demonstram eficientes na classificação. O modelo de RNA obteve resultados muito satisfatórios e por isso a decisão de trabalhar com seus parâmetros para tentar resultados ainda superiores do que já se encontrara.

Para trabalhos futuros, se sugere a investigação ainda mais profunda sobre a base de dados e outros algoritmos.

REFERÊNCIAS

[1] CHANG JUNIOR, Joao et al. Improving preoperative risk-of-death prediction in surgery congenital heart defects using artificial intelligence model: A pilot study. PLoS One, v. 15, n. 9, p. e0238199, 2020.

[2] VAID, Akhil et al. Multi-center retrospective cohort study applying deep learning to electrocardiograms to identify left heart valvular dysfunction. Communications Medicine, v. 3, n. 1, p. 24, 2023.

[3] DE SOUZA OLIVEIRA, Erivan et al. ABORDAGEM TERAPÊUTICA DOS PACIENTES PEDIÁTRICOS TRANSPLANTADOS CARDÍACOS. Saúde. com, v. 16, n. 2, 2020.

[4] ESTEVES, Hugo Filipe Padrão Brandão. Deteção de patologia cardíaca utilizando aprendizagem profunda. 2023.

[5] FERREIRA, Aparecido Pimentel; OLIVEIRA, Carlos ER; FRANÇA, Nancí Maria. Metabolic syndrome and risk factors for cardiovascular disease in obese children: the relationship with insulin resistance (HOMA-IR). Jornal de pediatria, v. 83, p. 21-26, 2007.