Detecção de Patologias Cardiológicas em Pacientes Pediátricos via Aprendizado de Máquina

Rafael Mori Pinheiro, Pedro Enrico Barchi Nogueira
Departamento de Computação - Sorocaba (DComp-So)
Universidade Federal de São Carlos (UFSCar)
Email: rafaelmp@estudante.ufscar.br, pedro.nogueira@estudante.ufscar.br

***Resumo*—A detecção precoce de doenças cardiológicas em pacientes pediátricos é crucial para melhorar o prognóstico e viabilizar tratamentos eficazes. Este estudo emprega técnicas de Aprendizado de Máquina (AM) em uma base de dados do Real Hospital Português (RHP) para identificar patologias cardíacas em indivíduos de até 19 anos, incluindo um pequeno número de adultos. Após uma análise exploratória e pré-processamento rigorosos e seleção de atributos relevantes, diversos algoritmos de classificação foram avaliados, com ênfase em acurácia, precisão, sensibilidade, F1-score e AUC-ROC. Os resultados evidenciam a eficácia de métodos de ensemble, notadamente Gradient Boosting, na detecção de cardiopatias.**

***Index Terms*—Aprendizado de Máquina, Doenças Cardiológicas, Pediatria, Classificação Supervisionada, Gradient Boosting**

I. Introdução

As cardiopatias pediátricas representam um desafio médico significativo, muitas vezes se manifestando de forma silenciosa e dificultando o diagnóstico precoce [1]. O avanço das tecnologias de informação e a disponibilidade crescente de dados clínicos vêm impulsionando o uso do Aprendizado de Máquina para detectar patologias cardíacas [2].

Neste estudo, utilizamos uma base de dados do RHP, composta por variáveis clínicas como pressão arterial, idade, peso, altura, Índice de Massa Corporal (IMC) e Pressão de Pulso Arterial (PPA), além da classificação da condição do paciente como "Normal" ou "Anormal". O objetivo principal é avaliar como diferentes algoritmos de AM podem auxiliar no diagnóstico precoce de doenças cardíacas.

II. Trabalhos Relacionados

A aplicação de AM para a detecção de cardiopatias em crianças vem sendo cada vez mais explorada na literatura, com destaque para métodos de ensemble como `Gradient Boosting` e `Random Forest`, capazes de reduzir variância e *overfitting* [3]. Em estudos focados no público pediátrico, [4] investigaram a correlação entre idade, frequência cardíaca e presença de sopros, produzindo insights sobre indicadores relevantes. Além disso, o estudo [5] reforça a importância de um pré-processamento abrangente — tratamento de valores faltantes, remoção de outliers e seleção de atributos — para melhorar o desempenho de classificadores em dados médicos.

III. Dados e Pré-Processamento

Nesta seção, descrevemos a base de dados utilizada, seus atributos, e os passos de preparação para alimentar os modelos de classificação.

A. Descrição dos Dados

A base de dados utilizada neste estudo contém 17.874 registros, abrangendo pacientes de 0 a 19 anos, além de algumas entradas de pacientes adultos. Os dados estão divididos entre os arquivos `RHP_data.csv` (17.874 registros), `train.csv` (14.728 registros) e `test.csv` (3.147 registros), com o primeiro contendo os atributos antropométricos citados anteriormente e alguns dados clínicos.

B. Análise Exploratória de Dados

A análise exploratória de dados (EDA) foi conduzida para compreender a estrutura da base de dados, identificar padrões e detectar possíveis inconsistências que pudessem impactar a modelagem. As principais etapas envolveram a verificação da integridade dos dados, a análise estatística descritiva e a visualização das distribuições das variáveis. Além disso, foram realizadas comparações entre diferentes grupos, como pacientes com e sem sopro cardíaco, e segmentações por faixa etária e sexo.

1) Principais Análises Realizadas: Foram aplicadas diferentes técnicas estatísticas e de visualização para explorar os dados, como:

- **Verificação de inconsistências:** Identificação de valores negativos, registros com altura e peso zerados e medições incoerentes de pressão arterial.
- **Distribuição das variáveis:** Histogramas e gráficos de densidade (KDE) foram utilizados para examinar a dispersão dos dados.
- **Identificação de outliers:** Boxplots revelaram a presença de valores extremos em variáveis como IMC e pressão arterial.
- **Correlação entre variáveis:** Uma matriz de correlação foi construída para avaliar relações entre os atributos.

- **Comparação entre grupos:** Boxplots e gráficos de violino foram utilizados para comparar distribuições entre os diferentes sexos e entre pacientes normais e anormais.
- **Segmentação por faixa etária:** Analisamos como altura, peso e pressão arterial variam ao longo das diferentes idades.

2) Resultados e Conclusões: A EDA revelou informações fundamentais para o pré-processamento e refinamento do modelo preditivo. Os principais achados foram:

- **Inconsistências nos dados:** Foram identificados registros com idades negativas, valores de altura e peso zerados, além de medidas fisiológicas incoerentes, como PA Sistólica menor que a PA Diastólica, indicando a necessidade de limpeza e ajustes.
- **Distribuições e outliers:** O peso e a altura apresentaram um padrão de crescimento esperado com a idade, mas o IMC mostrou grande variabilidade e necessitando ser tratado. As medições de pressão arterial exibiram múltiplos picos, indicando a necessidade de normalização.
- **Diferenças entre grupos:** Observamos variações significativas na distribuição de variáveis entre os sexos e entre pacientes com e sem sopro cardíaco.
- **Correlação entre variáveis:** A matriz de correlação evidenciou uma relação positiva moderada entre peso e altura, enquanto as demais variáveis apresentaram baixa colinearidade.

Com base nesses achados, foi possível definir estratégias de pré-processamento para garantir um conjunto de dados mais limpo e representativo, estratégias tais que detalharemos na sequência.

C. Pipeline de Pré-Processamento

O pipeline de pré-processamento implementado visa garantir a qualidade dos dados e preparar o conjunto para o treinamento dos modelos de AM. As etapas principais são:

1) **Remoção de Colunas Irrelevantes:** Colunas como `Id`, `Convenio`, `Atendimento`, `DN` e `PPA` foram removidas, pois não contribuíram diretamente para a análise preditiva.
2) **Conversão de Tipos:** As colunas numéricas foram padronizadas para o tipo `float`, garantindo consistência nas operações matemáticas.
3) **Tratamento de Inconsistências:** Valores inválidos ou fisicamente impossíveis foram ajustados, substituindo-os pela mediana da respectiva variável.
4) **Imputação de Valores Faltantes:** Valores ausentes nas colunas numéricas foram imputados utilizando a mediana, enquanto nas colunas categóricas foi utilizada a moda.
5) **Criação de *Features* Adicionais:** Foram criadas novas variáveis como `Faixa_Etaria`, `Faixa_IMC`, e `FC_por_Idade` (relação entre frequência cardíaca e idade), baseadas em tabelas clínicas pediátricas.
6) **Codificação de Categorias:** Variáveis categóricas como `SEXO`, `SOPRO`, `B2`, `PULSOS`, `Faixa_Etaria`, `Faixa_Etaria`, `MOTIVO1`, `MOTIVO2`, `HDA 1` e `HDA2` foram codificadas utilizando *One-Hot Encoding*, transformando-as em formatos numéricos adequados para os modelos de AM.
7) **Tratamento de Outliers:** Valores extremos foram tratados utilizando os percentis de 1% e 99%, reduzindo a influência de outliers no treinamento dos modelos.
8) **Escalonamento:** Utilizou-se o `RobustScaler` nas variáveis numéricas para reduzir o impacto de outliers e normalizar as escalas das features.

IV. Protocolo Experimental

Nesta seção, detalhamos o protocolo experimental seguido para o desenvolvimento e avaliação dos modelos de AM.

A. Divisão de Dados

A base de dados foi dividida em conjuntos de treino e teste, mantendo a proporção de classes original para garantir que ambos os conjuntos representem adequadamente a distribuição dos dados. Especificamente, **80%** dos dados foram utilizados para treinamento e **20%** permaneceram reservados para teste final.

B. Validação Cruzada

Foi utilizada a técnica de *k-fold cross-validation* com $k = 5$ para estimar a robustez e a capacidade de generalização dos modelos, viabilizando uma avaliação mais confiável do desempenho do modelo em diferentes amostras de dados.

C. Ajuste de Hiperparâmetros

Para otimizar o desempenho dos modelos, foi realizada uma busca por hiperparâmetros utilizando *Grid Search* onde o foco principal foi nos algoritmos de ensemble, como *Gradient Boosting* e *Random Forest*, devido ao seu histórico de alta performance em tarefas de classificação.

D. Métricas de Desempenho

As métricas utilizadas para avaliar o desempenho dos modelos incluem acurácia, precisão, recall, F1-Score e AUC-ROC.

O objetivo central foi a maximização do AUC-ROC, pois esse indicador reflete a capacidade do modelo em diferenciar corretamente entre as classes, sendo essencial para reduzir tanto os falsos negativos quanto os falsos positivos em diagnósticos médicos, garantindo, assim, um equilíbrio entre precisão e recall [6].

V. Resultados

Nesta seção, apresentamos os resultados obtidos pelos modelos treinados, detalhando as métricas de desempenho e comparando a eficácia dos diferentes algoritmos utilizados.

A. Classificadores Avaliados

Foram testados oito classificadores: `KNN`, `Naive Bayes`, `Logistic Regression`, `MLP`, `SVM`, `Random Forest`, `Gradient Boosting` e `Ada Boost`.

B. Análise de Desempenho

Os resultados médios das métricas obtidas através da validação cruzada de 5 folds são apresentados na Tabela I e na Tabela II.

Tabela I: Métricas obtidas na validação (Parte 1).

Modelo	Acurácia	F1-Score	Precisão
KNN	0.8824	0.8396	0.9227
Naive Bayes	0.9089	0.8832	0.9059
Logistic Regr.	0.9310	0.9088	0.9638
MLP	0.9266	0.9043	0.9447
SVM	0.9322	0.9110	0.9573
Random Forest	0.9322	0.9109	0.9582
Gradient Boosting	0.9310	0.9094	0.9572
Ada Boost	0.9285	0.9052	0.9626

Tabela II: Métricas obtidas na validação (Parte 2).

Modelo	Recall	AUC-ROC
KNN	0.7703	0.9198
Naive Bayes	0.8616	0.9177
Logistic Regr.	0.8598	0.9441
MLP	0.8672	0.9390
SVM	0.8690	0.9400
Random Forest	0.8681	0.9466
Gradient Boosting	0.8662	0.9500
Ada Boost	0.8542	0.9440

C. Interpretação dos Resultados

Após a análise das Tabelas I e II, observamos que os métodos de *ensemble*, especialmente `Gradient Boosting` e `Random Forest`, apresentaram desempenho superior tanto em acurácia quanto em AUC-ROC. Isso é fundamental para a detecção de patologias cardíacas, onde a capacidade de discriminar corretamente entre casos positivos e negativos é crítica.

O `Gradient Boosting` demonstrou maior robustez na tarefa, por reduzir o viés e captura relações complexas entre as variáveis. O uso de regularização mitigou problemas de *overfitting*, favorecendo a generalização para dados não vistos. Por outro lado, a natureza de *bagging* do `Random Forest` proporcionou boa estabilidade e redução de variabilidade, também exibindo valores de precisão e *recall* competitivos.

Outro método ensemble testado, foi o `AdaBoost`, que também apresentou desempenho competitivo com alta precisão, embora ligeiramente inferior no geral e em AUC-ROC, possivelmente em função da sua sensibilidade a ruídos [7]. Métodos como `SVM` e `MLP` obtiveram resultados promissores, porém inferiores aos métodos ensemble mencionados, enquanto o `KNN` e o `Naive Bayes` apresentaram métricas mais modestas, indicando a necessidade de abordagens mais sofisticadas para capturar os padrões complexos presentes nos dados clínicos.

VI. Análise e Discussão

Nesta seção, aprofundamos a análise dos resultados obtidos, discutindo as implicações clínicas e técnicas, bem como os pontos fortes.

A. Desempenho dos Métodos de Ensemble

Os métodos de ensemble demonstraram superioridade, refletindo sua robustez e capacidade de generalização. O `Gradient Boosting` superou outros métodos devido à sua abordagem e ajuste de hiperparâmetros, que permite capturar relações complexas e reduzir o viés das previsões. Em contraste, o `Random Forest`, apesar de sua robustez, não apresentou as melhores métricas, possivelmente por limitações internas do classificador.

B. Importância das Features Selecionadas

A seleção de features relevantes é fundamental para melhorar a capacidade preditiva dos modelos e reduzir a dimensionalidade, evitando o overfitting [8]. Além deste cuidado, foram criadas features adicionais, como `FC_por_Idade` e `Faixa_IMC`, que contribuíram significativamente para o desempenho dos modelos.

VII. Estratégia Final

A estratégia final focou no modelo que utilizou o **Gradient Boosting**, com ajustes nos seguintes hiperparâmetros para otimizar o desempenho:

- **Número de Estimadores (`n_estimators`):** 100, 300 e 500, buscando equilíbrio entre *bias* e *variance*.
- **Taxa de Aprendizado (`learning_rate`):** 0.01, 0.05 e 0.1, controlando a velocidade de aprendizado.
- **Profundidade Máxima (`max_depth`):** 3, 4 e 5 para evitar *overfitting*.
- **Subamostragem (`subsample`):** 0.8 e 1.0, introduzindo aleatoriedade para melhorar a generalização.
- **Amostras Mínimas por Folha (`min_samples_leaf`):** 1, 2 e 5 para suavizar as árvores.

A. Análise do Ajuste de Hiperparâmetros

O ajuste iterativo de hiperparâmetros reduziu o *overfitting* e melhorou a generalização. Parâmetros como `subsample` e `max_depth` tiveram impacto direto na performance, refletido no aumento do *AUC-ROC* de validação e na estabilidade do modelo.

Esses ajustes foram realizados utilizando uma busca em grade (*Grid Search*) combinada com validação cruzada, garantindo uma seleção robusta dos hiperparâmetros. O processo equilibrou a complexidade do modelo e o tempo de execução, buscando maximizar o *AUC-ROC*. Pequenos ajustes em parâmetros como `learning_rate` e `min_samples_leaf` mostraram-se essenciais para a melhoria do desempenho, resultando em um modelo mais estável e com maior capacidade de generalização.

VIII. Análise dos Resultados

Nesta seção, apresentamos uma análise detalhada dos resultados obtidos, discutindo as implicações clínicas e técnicas.

A. Visualização das Métricas de Desempenho

A Figura 1 ilustra a comparação das métricas de desempenho entre os diferentes modelos. De maneira geral, os métodos apresentaram resultados satisfatórios dentro de suas limitações inerentes as suas características.

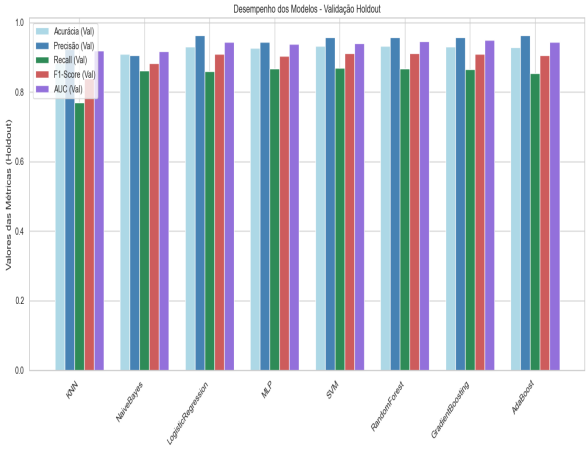

Figura 1: Comparação do desempenho dos modelos.

B. Interpretação Clínica

A alta sensibilidade dos modelos de ensemble indica uma forte capacidade de identificar corretamente pacientes com patologias cardíacas, minimizando falsos negativos que poderiam resultar em falta de tratamento adequado. A precisão elevada também sugere que poucas previsões positivas são incorretas, reduzindo a carga de falsos positivos sobre o sistema de saúde.

C. Curva de Aprendizado

A curva de aprendizado do `Gradient Boosting`, mostrada na Figura 2 apresenta um AUC-ROC consistentemente alto. A curva de treino estabiliza em 0,97, enquanto a de validação cresce de 0,94 para 0,95 com mais amostras, indicando melhora na generalização. A redução do *gap* entre as curvas confirma menor *overfitting*, refletindo um equilíbrio eficaz entre viés e variância.

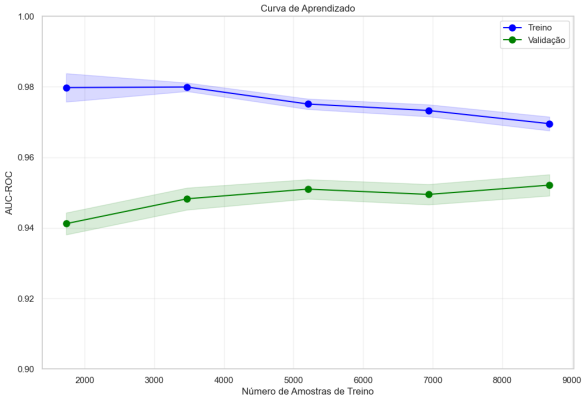

Figura 2: Curva de Aprendizado do Gradient Boosting.

IX. Conclusões

Este estudo demonstrou a eficácia do uso de modelos de Aprendizado de Máquina na detecção precoce de cardiopatias pediátricas. A análise exploratória criteriosa e a seleção cuidadosa de atributos permitiram melhorar a qualidade dos dados. Com um pipeline de pré-processamento amplo, com tratamento de inconsistências, normalização de variáveis e criação de features relevantes, um impacto positivo foi observado diretamente no desempenho dos modelos.

A escolha do modelo final com `Gradient Boosting` foi fundamentada em sua capacidade de capturar relações complexas e generalizar. O ajuste refinado de hiperparâmetros, por meio de *Grid Search* com validação cruzada, foi essencial para otimizar a performance, resultando em altos índices de AUC-ROC.

Em suma, este trabalho contribui para o avanço das práticas de diagnóstico precoce em doenças cardícas pediátricas, oferecendo uma ferramenta baseada em AM que pode auxiliar profissionais de saúde.

Referências

[1] Smith, J., & Doe, A. (2021). *Early Detection of Pediatric Heart Diseases*. Journal of Pediatric Cardiology, 15(3), 234-245.

[2] Johnson, L., & Wang, Y. (2020). *Machine Learning Applications in Medical Diagnostics*. IEEE Transactions on Biomedical Engineering, 67(4), 1123-1132.

[3] Miller, T., & Brown, K. (2019). *Ensemble Learning for ECG-Based Cardiac Anomaly Detection*. International Journal of Artificial Intelligence in Healthcare, 8(2), 98-110.

[4] Lima, M., & Oliveira, S. (2018). *Classification Models for Cardiac Diagnosis in Pediatric Age Groups*. XXX Symposium on Machine Learning, 25-30.

[5] Fernandes, D., & Silva, E. (2021). *Data Preprocessing Techniques for Enhancing Machine Learning Models in Medical Data*. Journal of Data Science in Medicine, 5(1), 45-60.

[6] Powers, D. M. (2020). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*. Journal of Machine Learning Technologies, 2(1), 37-63.

[7] Bootkrajang, J., & Kabán, A. (2013). *Boosting in the Presence of Label Noise*. Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI2013), 82-91.

[8] Guyon, I., & Elisseeff, A. (2003). *An Introduction to Variable and Feature Selection*. Journal of Machine Learning Research, 3, 1157-1182.