A detecção precoce de patologias cardiovasculares

Nome: Guilherme Fernandes Rezende Santos
RA: 813467

***Resumo* - A detecção precoce de patologias cardiovasculares é crucial na medicina pediátrica, uma vez que muitas dessas condições se desenvolvem de forma assintomática até estágios avançados. A identificação precoce pode melhorar significativamente o prognóstico e permitir tratamentos mais eficazes. Recentemente, avanços tecnológicos, incluindo o uso de aprendizado de máquina (AM), têm contribuído para aprimorar estratégias de diagnóstico, permitindo a análise eficiente de grandes volumes de dados médicos e auxiliando na tomada de decisão clínica. Este projeto utiliza uma base de dados real coletada no Real Hospital Português (RHP), no Brasil, referência em atendimento pediátrico e pioneiro na adoção de tecnologias de análise de dados para monitoramento cardiovascular.**

I. Introdução

A detecção precoce de patologias cardiovasculares em pacientes pediátricos é um tema de crescente importância na medicina, uma vez que essas condições podem se desenvolver de forma assintomática durante a infância e adolescência, muitas vezes só sendo identificadas em estágios avançados. As doenças cardiovasculares são a principal causa de morte no mundo, com aproximadamente 17,9 milhões de mortes em 2016, sendo que 17 milhões dessas mortes ocorreram de forma prematura, antes dos 70 anos de idade (OrganizaçãoPan-Americana de Saúde, 2016)[1].

A identificação precoce dessas patologias não apenas melhora o prognóstico dos pacientes, mas também possibilita a implementação de tratamentos preventivos mais eficazes, o que pode prevenir complicações graves e salvar vidas. A evolução das tecnologias de monitoramento e diagnóstico, aliada ao uso de ferramentas de aprendizado de máquina, tem revolucionado a forma como doenças cardiovasculares são diagnosticadas e tratadas em crianças e jovens. Em particular, o uso de dados clínicos coletados em ambientes hospitalares tem se mostrado uma estratégia poderosa para identificar padrões e fatores preditivos de doenças cardíacas em populações pediátricas.

A análise desses dados, que incluem informações como pressão arterial, índice de massa corporal, entre outros parâmetros vitais, oferece uma visão mais clara da saúde cardiovascular de crianças, permitindo a antecipação de tratamentos. Além disso, o uso de algoritmos de aprendizado de máquina para análise desses grandes volumes de dados tem mostrado resultados promissores na melhoria da precisão dos diagnósticos e na personalização dos cuidados médicos.

Esse cenário reflete uma mudança importante na abordagem tradicional do cuidado pediátrico, permitindo uma medicina mais proativa e personalizada, com o objetivo de salvar vidas e garantir uma melhor qualidade de vida para os pacientes em risco. A detecção precoce e o monitoramento contínuo das condições cardíacas desde a infância são passos fundamentais para combater o aumento das doenças cardiovasculares em adultos, que muitas vezes têm suas origens na infância.

A partir dessa perspectiva, o projeto visa desenvolver um sistema de detecção precoce de patologias cardiovasculares em pacientes pediátricos utilizando diferentes técnicas de aprendizado de máquina. O sistema será treinado e validado com a base de dados coletada no Real Hospital Português (RHP), visando identificar padrões e fatores de risco que possam indicar a presença de doenças cardíacas. A implementação desses modelos poderá auxiliar os profissionais de saúde na tomada de decisões clínicas mais informadas e na adoção de medidas preventivas, contribuindo para a melhoria da qualidade de vida e do prognóstico dos pacientes.

II. Trabalhos relacionados

O conjunto de dados fornecidos possui diversas colunas categóricas, o que torna necessário mapear esses dados para valores numéricos, afim de possibilitar o uso dos dados para treinamento do modelo. Em Aprendizado de Máquina para Predição de Diagnósticos de Doenças Cardiovasculares, Francisco Romes da Silva Filho e Emanuel F. Coutinho1.(2022), houve uma situação parecida, onde decidiram utilizar o multi-hot encoding para a tranformação dos dados, com base nessa escolha, assim, surgiu-se a ideia de aplicar o método One-hot-encoding para transformar os dados.

Segundo o estudo, as consequencias da DCV. Ghosh et al. (2021), é mostrado a eficácia do método de classificação Random Forest para a classificação de doenças cardíacas, o que abriu a possibilidade de usar o mesmo modelo nesse projeto.

[1]OPAN - Doenças cardiovasculares

III. Análise de dados e pré-processamento

O dataset foi desenvolvido pelo Real Hospital Português (RHP) com o objetivo de facilitar o estudo e a identificação de padrões associados a doenças cardiovasculares. O conjunto de dados é composto por um arquivo extenso no formato .csv, no qual cada linha corresponde a uma amostra distinta.

Cada amostra representa um paciente único que realizou uma consulta cardiológica e é caracterizada por 20 atributos, descritos a seguir:

- Peso do paciente em Kg.
- Altura do paciente em cm.
- IMC (índice de massa corporal) do paciente.
- Atendimento: data da consulta.
- DN: data de escrita da declaração de nascido vivo do paciente.
- Idade do paciente.
- Convênio do paciente.
- Pulsos: atributo categórico que indica a qualidade da circulação arterial do cliente.
- Pressão Arterial Sistólica (PA SISTOLICA): valor mais alto em mmHg que aparece durante uma aferição de pressão.
- Pressão Arterial Diastólica (PA DIASTOLICA): valor mais baixo em mmHg que aparece durante uma aferição de pressão.
- PPA (Pressão Pulso Arterial): atributo categórico que descreve o estado da pressão arterial de um paciente com base em medições clínicas.
- B2 (Segundo Ruído Cardíaco): atributo categórico que representa o estado do som de fechamento das válvulas aórticas e pulmonar.
- Sopro: atributo categórico que está relacionado à ausculta cardíaca e descreve a presença e características de sopros no coração.
- Frequência Cardíaca (FC) do paciente medidos em batimentos por minuto (bpm).
- HDA1 (Histórico de doenças atual 1): representa o primeiro problema clínico do histórico do paciente.
- HDA2 (Histórico de doenças atual 2): representa o segundo problema clínico do histórico do paciente.
- Genêro biológico do paciente.
- Motivo 1: primeira razão para a consulta.
- Motivo 2: segunda razão para a consulta.

A. Preparação dos dados

O conjunto de dados contém dados de treino e dados de teste juntos, sendo cada um deles identificado pelo Id da amostra. Como os dados de teste não possuem a classe da amostra, não podem ser utilizados durante o processo de treinamento dos modelos. Levando isso em conta, separamos os dados do dataset RHP.csv em dois outros sub-conjuntos:

- train-data: contém as amostras utilizadas para treinamento dos modelos.
- test-data: contém as amostras utilizadas para as predições no Kaggle.

B. Análise dos dados

A análise dos dados foi conduzida por meio da visualização e interpretação de gráficos, os quais relacionam os atributos com a classe alvo. Para os atributos categóricos, foram utilizados gráficos de barras, que permitem visualizar a distribuição das diferentes categorias de cada atributo em relação à classe. No caso dos atributos numéricos, optou-se por gráficos boxplot, que possibilitam a análise da distribuição dos valores e sua dispersão em relação à classe.

Após a análise dos gráficos, pode-se obter um planejamento do que deveria ser realizado no pré-processamento, identificando a necessidade de transformações específicas nos dados. Por exemplo, a presença de valores irreais em atributos numéricos pode demandar imputação, enquanto a existência de valores redundantes ou inconsistentes em colunas categóricas pode requerer a padronização desses valores. Além disso, para atributos categóricos com múltiplas categorias, é essencial aplicar técnicas como o one-hot encoding.

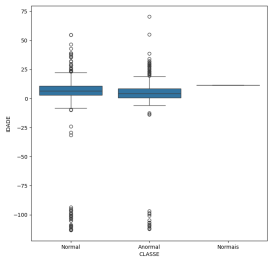

Figure 1. Box plot da idade

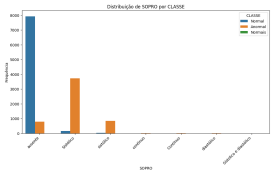

Figure 2. Gráfico de sopro

O atributo classe foi analisado por meio de um gráfico de pizza, que ilustra a distribuição das classes no conjunto de treinamento. A partir da visualização, observa-se que a proporção de amostras da classe Normal excede a da classe Anormal em 20%. Essa disparidade pode levar a um viés no modelo, favorecendo a classe majoritária. Logo, é evidente a necessidade de remover uma certa quantidade de amostras pertencentes à classe Normal, visando evitar o viés.

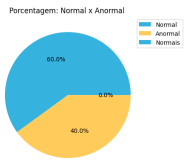

Figure 3. distribuicao de classes no conjunto de treino