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.

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, a fim 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. Coutinho.(2022)[1], houve uma situação semelhante, na qual foi utilizado o método multi-hot encoding para a transformação dos dados. Com base nessa escolha, surgiu a ideia de aplicar o método one-hot encoding para transformar os dados.

Segundo o estudo sobre as consequências da DCV.Ghosh et al. (2021) [2], foi demonstrada a eficácia do método de classificação Random Forest para a classificação de doenças cardíacas, o que abriu a possibilidade de utilizar o mesmo modelo neste projeto.

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.

[1]OPAN - Doenças cardiovasculares

- 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.

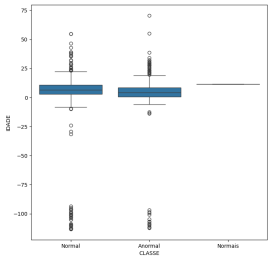

Figure 1. Box plot da idade

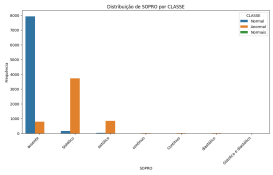

Figure 2. Gráfico de sopro

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.

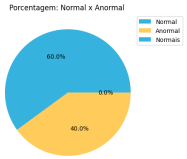

Figure 3. distribuicao de classes no conjunto de treino

C. *Pré-processamento*

Devido ao número limitado de amostras no conjunto de treino, optou-se por preservar o máximo possível de dados, preferindo a imputação de valores nulos e irregulares nas colunas numéricas em vez de sua remoção. Primeiro, valores irreais foram definidos como nulos para não distorcer as medidas descritivas. Em seguida, aplicou-se a imputação pela mediana nos valores faltantes, após isso , o IMC foi recalculado com base nos novos valores. A mediana foi escolhida devido ao alto desvio padrão em relação à grandeza dos atributos.

Peso	Altura	IDADE	PA SISTOLICA	PA DIASTOLICA	FC
17.31	33.29	4.59	10.58	8.68	16.63

Valores do desvio padrão de cada atributo tratado.

Para as colunas categóricas, a maioria das amostras com valores faltantes foi removida, exceto em "Pulsos", "HDA2", "MOTIVO2" e "PPA". Os três primeiros casos foram mantidos devido ao grande número de valores faltantes, cuja remoção prejudicaria o treinamento do modelo, enquanto "PPA" continha a categoria 'Não calculado', que foi usada para substituir os nulos.

No conjunto de testes, a abordagem foi similar, com a diferença de que todos os valores faltantes nas colunas categóricas foram substituídos pela moda, já que a remoção de amostras não era permitida.

Após o tratamento dos valores espúrios, criou-se o atributo 'SOPRO PESO', combinando 'SOPRO' e 'Peso'. A decisão foi motivada pela alta correlação do atributo target com ambos, além da forte correlação entre eles. Para evitar redundância, o atributo 'Peso' foi removido, já que sua informação foi incorporada no novo atributo.

As colunas numéricas foram normalizadas utilizando o método QuantileTransformer. Essa abordagem teve como objetivo reduzir o impacto de outliers no modelo e padronizar a escala dos valores, já que alguns algoritmos, como a Regressão Logística, são sensíveis a diferenças de grandeza entre os atributos.

As colunas categóricas foram mapeadas através do método One Hot Label Encoding, para evitar a criação de pesos ou hierarquias inadequadas aos valores.

IV. Protocolo experimental

Neste projeto, foram implementados os seguintes modelos de aprendizado: K-Neighbors, Naive Bayes, Regressão Logística, Redes neurais artificiais, Máquinas de vetores de suporte e floresta aleatória. A técnica de florestas aleatórias foi escolhida pois a mesma mostrou bons resultados para diversos trabalhos relacionadoa à dados médicos.

Como modelos como florestas aleatórias, máquinas de vetores de suporte, redes neurais artificiais e regressão logística são sensíveis à alteração de seus hiper-parâmetros, foi utiliziado o método de GridSearch para encontrar os melhores parâmetros para cada um desses modelos. A busca foi feita utilizando o conjunto de treino utilizando cross-validation com 5 folds, onde a combinação que apresentou a melhor acurácia seria selecionada. A busca foi estruturada da seguinte forma:

- Na regressão logística foram testados valores de C em um intervalo de 10^{-2} até 10^{3} com passo de 1 na potência.
- O número de hidden-layer-sizes para redes neurais artificiais foi testado em um intervalo de 10 até 40.
- No caso de máquinas de vetores de suporte com kernel linear, foram testados valores de C em um intervalo de 10^{-2} até 10^{3} com passo de 1 na potência.
- Para máquinas de vetores de suporte com kernel linear e com kernel polinomial, foram testados valores de C e γ em um intervalo de 10^{-2} até 10^{3} com passo de 1 na potência.
- O número de estimators em random forest foi testado em um intervalo de 100 a 300, com incremento de 50.

Após o término da busca, os seguintes valores de hiperparâmetros foram escolhidos:

- Regressão Logística: $C = 1$.
- Florestas aleatórias: 150 estimators.
- Redes Neurais Artificiais: 10 hidden layers.
- Máquinas de vetores de suporte (linear): $C = 0.1$.
- Máquinas de vetores de suporte (radial): $C = 1, \gamma = 0.1$.
- Máquinas de vetores de suporte (polinomial): $C = x, \gamma = x$.

Todos os modelos, assim como os métodos de pré-processamento e avaliação, foram implementados utilizando a biblioteca scikit-learn.

Afim de obter os resultados avaliativos, foi utilizado o método cross-validation com 5 folds. Para avaliar os modelos, foram calculadas a média e desvio padrão de cada uma das seguintes medidas de avaliação:

- Acurácia - Proporção de previsões corretas em relação ao total de previsões.
- Precisão - Proporção de verdadeiros positivos entre todas as previsões classificadas como positivas.
- Recall - Proporção de verdadeiros positivos identificados corretamente em relação a todos os casos positivos reais.
- F1 Score - Média harmônica entre precisão e recall, equilibrando os dois indicadores.

Adicionalmente, a curva ROC foi plotada a partir da divisão do conjunto de treinamento em 80% para treino e 20% para teste. Essa análise visou identificar o modelo com o maior valor de AUC, permitindo selecionar a melhor opção para a competição com base no desempenho discriminativo.

V. Resultados

Modelo	Acurácia (%)	Precisão (%)	Recall (%)	F1 Score (%)
Random Forest	88.63 ± 1.74	72.28 ± 5.45	3.20 ± 1.58	0.86 ± 0.02
SVM Radial	87.83 ± 2.10	72.17 ± 6.17	4.34 ± 1.59	0.85 ± 0.03
SVM Polinomial	87.45 ± 1.92	71.27 ± 5.72	4.45 ± 1.52	0.85 ± 0.02
SVM Linear	87.33 ± 2.12	70.51 ± 5.74	4.25 ± 1.84	0.85 ± 0.02
Regressão Logística	86.10 ± 1.98	71.24 ± 5.27	6.46 ± 1.96	0.84 ± 0.02
Redes Neurais	86.10 ± 1.98	71.24 ± 5.27	6.46 ± 1.96	0.84 ± 0.02
7-NN	85.67 ± 1.99	76.96 ± 5.11	5.73 ± 2.17	0.83 ± 0.03
5-NN	85.52 ± 1.72	70.84 ± 4.77	6.96 ± 2.17	0.83 ± 0.02
3-NN	84.07 ± 2.87	71.75 ± 5.53	9.77 ± 3.10	0.82 ± 0.03
Naïve Bayes	82.51 ± 2.15	72.06 ± 4.88	12.27 ± 2.42	0.80 ± 0.03

A tabela acima apresenta a média e desvio padrão de todas as medidas de desempenho selecionados, em ordem decrescente de acurácia.

A análise dos resultados demonstra que a maioria dos métodos avaliados foi eficaz na detecção de pacientes com condições cardíacas anormais, apresentando desempenho satisfatório em termos de precisão e recall. Observa-se que as técnicas de classificação conseguiram identificar corretamente uma proporção significativa de casos anormais, com destaque para métodos como Random Forest e máquinas de vetores de suporte (SVM) com diferentes kernels, que alcançaram um equilíbrio adequado entre a detecção de casos anormais e a minimização de falsos positivos.

No entanto, é crucial ressaltar que alguns métodos apresentaram taxas elevadas de classificação incorreta de pacientes saudáveis como anormais, o que pode comprometer sua aplicação em cenários clínicos reais. Nesse aspecto, técnicas como Random Forest e SVM com kernel polinomial se destacaram por apresentarem as menores taxas de falsos positivos, sendo consideradas mais adequadas para aplicações práticas.

Em termos de desempenho geral, o Random Forest obteve os melhores resultados, com um equilíbrio notável entre acurácia, precisão e recall. Métodos como regressão logística e SVM com kernels radial e linear também demonstraram resultados consistentes, embora ligeiramente inferiores. Por outro lado, técnicas baseadas em k-NN e Naive Bayes apresentaram desempenho menos satisfatório, principalmente devido às taxas mais altas de classificação incorreta de pacientes saudáveis.

Os desempenhos obtidos no public leaderboard da competição do Kaggle demonstraram consistência e robustez, com as acurácias variando entre 93,7% e 93,8%. Esses resultados refletem não apenas a alta capacidade de generalização dos modelos desenvolvidos, mas também a eficácia das técnicas empregadas na tarefa de classificação. A proximidade entre as métricas alcançadas indica que os modelos mantiveram um equilíbrio entre precisão e recall, garantindo uma detecção confiável de casos anormais sem comprometer significativamente a taxa de falsos positivos.

VI. Estratégia final

Para determinar quais modelos seriam utilizados na competição, realizou-se a plotagem das curvas ROC e o cálculo dos respectivos valores de AUC (Área Sob a Curva), utilizando um conjunto de 80% treino e 20% teste. Como a métrica de avaliação da competição era baseada no valor de AUC, optou-se por selecionar os dois modelos que apresentaram os maiores valores dessa métrica, garantindo assim um desempenho superior na classificação. Após a análise das curvas ROC e dos cálculos, os valores de AUC obtidos para cada modelo foram organizados e estão apresentados na tabela abaixo.

Modelo	AUC
Random Forest	95.98%
Regressão Logística	95.51%
Redes Neurais	95.42%
SVM radial	95.11%
7-NN	94.64%
SVM polinomial	94.56%
5-NN	94.36%
3-NN	94.36%
Naive Bayes	94.16%
SVM linear	93.28%

AUC de cada modelo

Com base na análise das curvas ROC e nos valores de AUC, os modelos selecionados para a competição foram o Random Forest, com 150 estimators, e a Regressão Logística, com o parâmetro de regularização C=1. Esses modelos se destacaram por apresentarem os maiores valores de AUC, demonstrando alta capacidade de distinção entre as classes e um desempenho superior na tarefa de classificação.

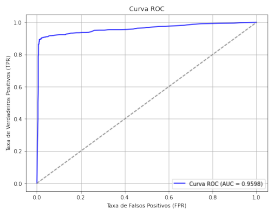

Figure 4. Curva ROC do modelo random forest

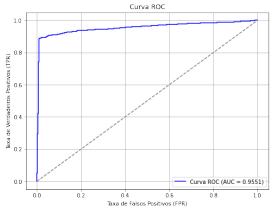

Figure 5. Curva ROC do modelo regressão logística

VII. Conclusões

Este trabalho teve como objetivo desenvolver um sistema para a detecção precoce de patologias cardiovasculares em pacientes pediátricos, empregando diversas técnicas de aprendizado de máquina.

Os resultados obtidos demonstram que os métodos de classificação avaliados foram eficazes na detecção de pacientes com condições cardíacas anormais, garantindo um bom equilíbrio entre precisão e recall. Em especial, o Random Forest se destacou por apresentar um desempenho robusto, conciliando alta acurácia com uma baixa taxa de falsos positivos. Modelos como SVM com kernel polinomial e regressão logística também mostraram resultados consistentes, reforçando sua viabilidade para aplicações clínicas.

No entanto, algumas técnicas apresentaram limitações, como o k-NN e o Naive Bayes, que tiveram dificuldades na correta classificação de pacientes saudáveis, resultando em um número elevado de falsos positivos. Esse fator ressalta a importância da escolha criteriosa do modelo para garantir um diagnóstico mais confiável.

Como sugestão de extensão para o trabalho, é interessante a ampliação do conjunto de dados por meio da inclusão de atributos clínicos mais detalhados, especialmente aqueles diretamente relacionados à saúde cardiovascular.

REFERENCES

[1] Aprendizado de Máquina para Predição de Diagnósticos de Doenças Cardiovasculares, Fran- cisco Romes da Silva Filho e Emanuel F. Coutinho.(2022)

[2] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shul- tana, S., Beeravolu, A. R., and De Boer, F. (2021). Efficient prediction of cardiovas- cular disease using machine learning algorithms with relief and lasso feature selection techniques.