Comparação de Algoritmos de Aprendizado de Máquina para previsão de doenças cardíacas

Arthur Crossy Reis Mendes
PUC Minas
Belo Horizonte, Brasil

Arthur Gonçalves de Moraes
PUC Minas
Belo Horizonte, Brasil

Davi Martins Freitas Wanderley
PUC Minas
Belo Horizonte, Brasil

Gabriel Araujo Campos Silva
PUC Minas
Belo Horizonte, Brasil

Thiago Cedro Silva de Souza
PUC Minas
Belo Horizonte, Brasil

RESUMO

Este estudo apresenta uma análise detalhada da comparação de vários algoritmos de aprendizado de máquina para a previsão de doenças cardiacas, utilizando dados extraídos da base de dados Kaggle. Iniciamos com um processo de pré-processamento de dados, que incluiu a substituição de valores ausentes,removoção de duplicidades para mitigar desequilíbrios entre classes. Além disso, implementamos técnicas para a normalização dos dados, a fim de prepará-los eficazmente para análise. Testamos algoritmos de aprendizado de máquina: Random Forest Naive Bayes e Decision Tree. Os resultados obtidos demonstram que o modelo Random Forest se destacou na classificação dos indivíduos com relação ao diagnóstico de doença cardíacas, refletindo a importância de um processo de pré-processamento de dados. Este trabalho evidencia a importância do potencial do uso de algoritmos de aprendizado de máquina na predição de condições de saúde.

KEYWORDS

Aprendizado de Máquina, Random Forest, Kaggle, Previsão de Saúde, Saúde Pública, Análise de Dados

ACM Reference Format:
Arthur Crossy Reis Mendes, Arthur Gonçalves de Moraes, Davi Martins Freitas Wanderley, Gabriel Araujo Campos Silva, and Thiago Cedro Silva de Souza. 2024. Comparação de Algoritmos de Aprendizado de Máquina para previsão de doenças cardíacas . In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

As doenças cardíacas é reconhecida globalmente como um dos principais motivos de mortalidade de pessoas adultas e idosas [3]. A detecção precoce e o gerenciamento adequado da doenças cardíacas são cruciais para prevenir complicações a longo prazo[3]. No entanto, muitos indivíduos com doenças cardíacas permanecem não diagnosticados ou não tratados, em parte devido à natureza assintomática da condição nas fases iniciais [3].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

A utilização de algoritmos de aprendizado de máquina na área da saúde tem demonstrado grande potencial para melhorar a precisão e a eficiência do diagnóstico e da previsão de doenças [**?**]. Entre esses algoritmos, destacam-se o Random Forest, Naive Bayes e Decision Tree, cada um com suas vantagens específicas para lidar com grandes volumes de dados e variáveis preditoras [**?**]. Este trabalho concentra-se na aplicação desses algoritmos para prever a ocorrência de doenças cardíacas, utilizando o conjunto de dados da Kaggle [1]. A escolha deste conjunto de dados é motivada pela necessidade de entender melhor os fatores de risco associados á doenças cardíacas[3]. Além disso, o estudo busca explorar o impacto de um conjunto selecionado de variáveis relacionadas ao estilo de vida, condições socioeconômicas e demográficas na prevenção de doenças cardíacas [1].

1.1 Coleta e Pré-processamento dos Dados

Os dados analisados neste estudo foram obtidos da base de dados Kaggle, que inclui informações detalhadas sobre uma amostra de pessoas[1].

1.1.1 Substituição de Valores Ausentes. Valores ausentes foram substituídos pela mediana das colunas respectivas para evitar viés significativo [4]. Esta abordagem foi escolhida porque a imputação pela mediana é uma técnica simples e eficaz que mantém a distribuição dos dados e minimiza a possiblidade de outliers [4]. Alternativas como a imputação por médias ou modos foram consideradas, mas a medianas foi selecionada para reduzir a possibildaide de outliers [4].

1.1.2 Duplicidade. As duplicidades foram removidas para manter a correta distribuição dos dados e minimizar a possibilidade de desvios da médias.

1.1.3 Codificação. a maioria dos atributos fosse numérica. Entretanto, quando existem atributos categórico realizou - se uma transformação para valores numéricos utilizando One-Hot Encoding, facilitando a aplicação dos algoritmos de aprendizado de máquina [2]. Dessa forma, A One-Hot Encoding transforma atributos categóricos em vetores binários, com isso evita a introdução de ordens aleatórias que podem distorcer os resultados dos modelos preditivos [2].

Este processo de pré-processamento garantiu que os dados estivessem em uma forma adequada para a análise e modelagem subsequente, permitindo a aplicação eficaz dos algoritmos de aprendizado de máquina [2].

1.2 Seleção de Atributos

Os atributos selecionados foram escolhidos com base em estudos anteriores [3] e relevância teórica para a previsão de doenças cardíacas [3]. Cada atributo foi incluído no modelo devido à sua associação comprovada com o risco de doenças cardíacas [1] ou seu papel protetivo [1].

A seguir, apresentamos uma descrição detalhada dos principais atributos e suas importâncias, juntamente com referências aos estudos que suportam sua inclusão:

- **Age**: Idade do paciente em anos. [1].
- **Sex**: Atributo binário, cujo é um indicador do sexo do indivíduo. Estudos indicam que fatores de risco e prevalência de doenças cardíacas podem diferir entre homens e mulheres, nesse sentido, podemos visualizar o impacto do genêro de doenças cardiácas. [1].
- **cp - Tipos de dores no peito (1-4)**: Atritibuto continúo, cujo é um tipo de dores no peito, ou seja, as dores no peito estão relacionados com os sintomas de uma possível doença cardíaca. Nesse sentido, evidência - se a necessidade de selecionar esse atributo para uma melhor analíse de dados.[1], [3].
- **trestbps (Pressão arterial em repouso)**: Atritibuto discreto, cujo é um indicador do nível . A pressão arterial em repouso é medida quando a pessoa está relaxada e em um ambiente tranquilo. É considerada um dos principais indicadores de saúde cardiovascular, pois permite avaliar a eficiência do coração em bombear sangue para o corpo e a condição das artérias. [1].
- **chol (Colesterol sérico em mg/dl)**: Atritibuto discreto, cujo é um indicador de colesterol LDL, ou colesterol "ruim", está associado a um maior risco de problemas como infarto ou AVC quando elevado. [1].
- **fbs (Açúcar no sangue em jejum)**: Atritibuto binário, o estado de normalidade da glicemia em jejum [1].
- **restecg (Resultados eletrocardiográficos em repouso)**: Atritibuto binário, Um ECG com laudo normal indica que a atividade elétrica do coração está em pleno funcionamento. Relata ainda que não existe grandes entupimentos das coronárias e que não há aumento de cavidades ou arritmias. [1].
- **thalach (Frequência cardíaca máxima alcançada)**: frequência cardíaca é um atributo discreto. Além do mais, ela representa a velocidade do ciclo cardíaco medida pelo número de contrações do coração por minuto (bpm). [1].
- **exang (Angina induzida por exercício)**: Atritibuto binário, Angina é uma dor no peito temporária ou uma sensação de pressão que ocorre quando o músculo cardíaco não está recebendo oxigênio suficiente. O indivíduo com angina costuma sentir desconforto ou pressão abaixo do esterno.[1].
- **oldpeak (Depressão do segmento ST induzida por exercício em relação ao repouso)**: Atritibuto binário, ademais quando um indivíduo realiza atividade física, a demanda de oxigênio para o coração aumenta. Se as artérias coronárias, responsáveis por fornecer sangue ao coração, estiverem estreitadas ou obstruídas (como ocorre na doença arterial coronariana), o fluxo sanguíneo pode não ser suficiente para atender essa demanda. Isso pode levar a uma depressão do segmento ST no ECG.
- **slope (Inclinação do Segmento ST no Pico de Exercício)**: Descreve a inclinação do segmento ST no pico do exercício (0 = inclinado para cima, 1 = plano, 2 = inclinado para baixo).
- **ca (Número de Vasos Principais)**: Número de vasos principais (variando de 0 a 3) visíveis por fluoroscopia.
- **thal (Talassemia)**: Um distúrbio sanguíneo (1 = normal, 2 = defeito fixo, 3 = defeito reversível).
- **Target (Alvo)**: Indica a presença ou ausência de doença cardíaca (1 = presença, 0 = ausência).
  [1].

A importância relativa desses atributos foi avaliada utilizando o modelo Random Forest, que permite a identificação dos atributos mais relevantes para a previsão de doenças cardíacas. A Figura **??** ilustra a importância de cada atributo no modelo [**?**].

1.3 Modelagem

Para a modelagem, utilizou-se uma variedade de algoritmos de aprendizado de máquina, incluindo Random Forest, Naive Bayes e Decision Tree, devido à sua capacidade de manejar grandes bases de dados e lidar com a não-linearidade entre os atributos e o diagnóstico de doenças cardíacas [**?**].

1.4 Avaliação do Modelo

A avaliação dos modelos foi realizada utilizando diversas métricas de desempenho, garantindo uma análise abrangente da eficácia preditiva de cada algoritmo. As métricas utilizadas foram:

- **Acurácia**: A proporção de previsões corretas, calculada como o número total de previsões corretas dividido pelo número total de amostras. Esta métrica fornece uma visão geral da capacidade do modelo de fazer previsões corretas [**?**].
- **Precisão**: A proporção de verdadeiros positivos entre as previsões positivas, calculada como $\frac{\text{VP}}{\text{VP+FP}}$. A precisão é importante em contextos onde o custo de falsos positivos é alto [**?**].
- **Recall (Sensibilidade)**: A proporção de verdadeiros positivos identificados corretamente, calculada como $\frac{\text{VP}}{\text{VP+FN}}$. O recall é crucial em contextos onde é importante identificar todos os casos positivos, mesmo que alguns negativos sejam classificados erroneamente [**?**].
- **F1-score**: A média harmônica entre precisão e recall, calculada como $2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão+Recall}}$. O F1-score é uma métrica equilibrada que considera tanto a precisão quanto o recall, sendo útil quando há um trade-off entre essas duas métricas [**?**].

A utilização dessas métricas permite uma avaliação extensa dos modelos, considerando diferentes aspectos do desempenho dos algoritmos. [**?**].

2 RESULTADOS

2.1 Comparação de Desempenho dos Algoritmos

Os resultados avaliados a partir dos diferentes algoritmos de aprendizado de máquina são apresentados na Tabela 1. Dessa forma, a

tabela mostra as métricas de acurácia, precisão, recall e F1-score para cada modelo.

Modelo	Acurácia	Precisão	Recall	F1-score
Random Forest	0.82	0.88	0.85	0.87
Naive Bayes	0.84	0.84	0.91	0.87
Decision Tree	0.836	0.83	0.71	0.76

Tabela 1: Comparação de desempenho dos diferentes algoritmos.

Além das tabelas, os gráficos a seguir ilustram as métricas de desempenho de cada modelo, facilitando a comparação visual.

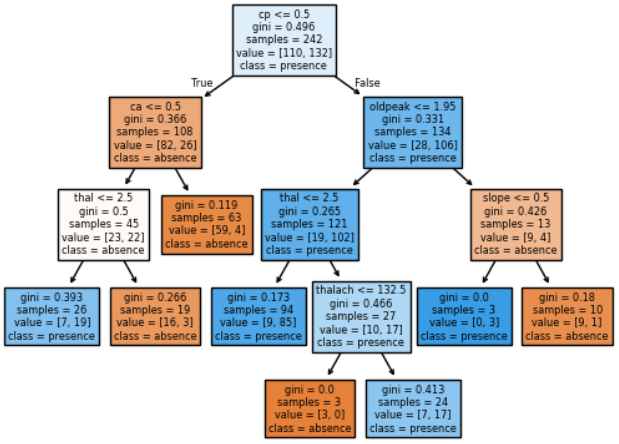

Figura 1: Árvore de decisão.

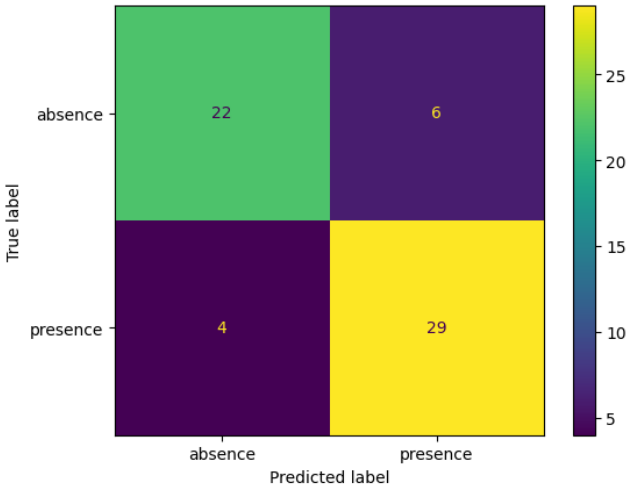

Figura 2: Matriz de confusão.

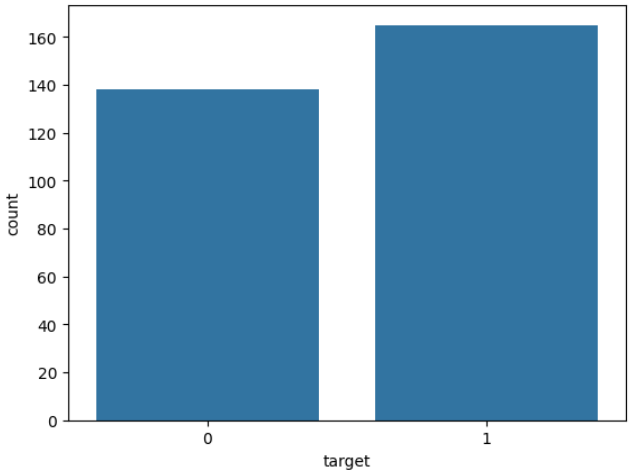

Figura 3: Quantidade de instâncias do target

2.2 Importância das Características no Random Forest

A figura **??** mostra os atributos no modelo Random Forest, destacando quais variáveis têm maior impacto na previsão de doenças cardíacas.

2.3 Análise da Matriz de Confusão

As matrizes de confusão para o melhor modelo (Random Forest) e o pior modelo Decision Three) são apresentadas a seguir. Essas matrizes mostram a capacidade de cada modelo em identificar corretamente os casos de doenças cardíacas, cujos quais, reporesentam os (verdadeiros positivos) e os casos de não-doenças negativas, ou seja, os (verdadeiros negativos) [**?**].

2.4 Árvore de Decisão

A figura 1 apresenta a árvore de decisão gerada pelo modelo Decision Tree, destacando os principais fatores associados às doenças cardíacas.

3 DISCUSSÃO

3.1 Comparação com a Literatura

Os resultados deste estudo são relativamente consistentes com pesquisas anteriores que destacam a eficácia do Random Forest na predição de condições de saúde, incluindo doenças cardíacas. No entanto, ele apresenta uma acurácia menor em relação ao outros algoritmos. Trabalhos como o de [**?**] demonstraram que o Random Forest pode superar outros algoritmos em termos de precisão e robustez. [**?**].

Além disso, a apesar do Naive Bayes apresentar - se um pouco melhor em relação ao Random Forest. Observa - se neste estudo, cujo qual, contradiz levemente a pesquisa [**?**], que sugere que a suposição de independência condicional dos atributos pode limitar a eficácia deste algoritmo em conjuntos de dados com características altamente conectados, como no campo da saúde.

4 CONCLUSÃO

Apesar dos resultados promissores, existem inúmeras formas de aprimorar esse estudo, como a criação

5 UTILIZAÇÃO DO GPT

Durante o desenvolvimento deste trabalho, a ferramenta GPT (Generative Pre-trained Transformer) foi utilizada de forma restrita para auxiliar algumas tarefas básicas de formatação. Dessa forma, houve a utilização para verificação de consistência textual e geração de códigos para plotagem de gráficos. Com isso, as áreas específicas onde o GPT foi empregado incluem:

- Correção ortográfica e gramatical de trechos do texto.
- Sugestões de formatação em LaTeX.
- Auxílio na elaboração de seções introdutórias e conclusivas.
- Geração de códigos em Python.

É importante destacar que todas as análises de dados, implementação de algoritmos de aprendizado de máquina, interpretação dos resultados e discussões foram realizadas integralmente pelos autores. O uso do GPT foi limitado a suportes textuais, de formatação e de visualização, não influenciando no conteúdo técnico e científico do trabalho.

6 CÓDIGO DESENVOLVIDO

O código desenvolvido pode ser acessado no seguinte link: https://colab.research.google.com/drive/1uqF2o35Mm9dL2cfUC1Dm3hbWtowNXmZC?usp=sharing.

REFERÊNCIAS

[1] Base de dados utilizada. page 0, 2021.
[2] Taher Al-Shehari and Rakan A Alsowail. An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, 23(10):1258, 2021.
[3] Alanna Gomes da Silva Ana Carolina Micheletti Gomide Nogueira de Sá Francielle Thalita Almeida Alves Antonio Luiz Pinho Ribeiro Deborah Carvalho Malta journal=Revista Brasileira de Epidemiologia year=2019 Crizian Saar Gomes, Renata Patrícia Fonseca Gonçalves. Fatores associados às doenças cardiovasculares na população adulta brasileira: Pesquisa nacional de saúde, 2019.
[4] Babu Sena Paul Luke Oluwaseye Joel, Wesley Doorsamy. A review of missing data handling techniques for machine learning. *University of Johannesburg,*, 23(10):1258, 2021.