

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

Arthur Crossy Reis Mendes
PUC Minas
Belo Horizonte, Minas Gerais

Arthur Gonçalves de Moraes
PUC Minas
Belo Horizonte, Minas Gerais

Davi Martins Freitas Wanderley
PUC Minas
Belo Horizonte, Minas Gerais

Gabriel Araujo Campos Silva
PUC Minas
Belo Horizonte, Minas Gerais

Thiago Cedro Silva de Souza
PUC Minas
Belo Horizonte, Minas Gerais

RESUMO

Este estudo apresenta uma análise detalhada da comparação de vários algoritmos de aprendizado de máquina para a previsão de doenças cardíacas, utilizando dados extraídos da base de dados Kaggle. Iniciamos com um processo de pré-processamento de dados, que incluiu a substituição de valores ausentes, remoção 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: CART (Decision Tree), Backpropagation (Redes Neurais) e KNN. Os resultados obtidos demonstram que o modelo KNN se destacou na classificação dos indivíduos com relação ao diagnóstico de doença cardíacas, refletindo a importância de um processo de pré-processamento de dados.

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 *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/1122445.1122456>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2024 Association for Computing Machinery.

ACM ISBN 978-1-4503-9999-9/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

1 INTRODUÇÃO

As doenças cardíacas são reconhecidas globalmente como um dos principais motivos de mortalidade de pessoas adultas e idosas. A detecção precoce e o gerenciamento adequado da doenças cardíacas são cruciais para prevenir complicações a longo prazo. No entanto, muitos indivíduos com doenças cardíacas permanecem não diagnosticados ou não tratados, em parte devido à natureza assintomática da condição nas fases iniciais. 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 Decision Tree, Backpropagation e KNN, 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. A escolha deste conjunto de dados é motivada pela necessidade de entender melhor os fatores de risco associados a doenças cardíacas. 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.

2 OBJETIVOS

A área da saúde está em constante evolução e anualmente novas tecnologias de outras áreas são incorporadas no dia-a-dia dos profissionais da área, como robótica e computação. Devido a isso, encontrou-se um espaço para estudar formas de aplicar o Aprendizado de Máquina no meio da saúde. Ao pesquisar a respeito, percebeu-se uma vasta gama de pesquisas sobre o assunto, porém a maioria com os mesmos modelos sendo analisados. Portanto, esse trabalho tem o objetivo de explorar novas opções de métodos de aprendizado de máquina, além de outras configurações para métodos mais abordados, aumentando o conhecimento da população acadêmica de ambas as áreas e estreitando a comunicação sobre os modelos de IA que funcionam para diagnóstica doenças.

3 COLETA E PRÉ-PROCESSAMENTO DE DADOS

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

3.1 Substituição de Valores Ausentes

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

3.2 Duplicidade

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

3.3 Codificação

Apesar de que a maioria dos atributos fosse numérica, como existem atributos categóricos realizou - se uma transformação para valores numéricos utilizando One-Hot Encoding, facilitando a aplicação dos algoritmos de aprendizado de máquina. 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.

3.4 Normalização

Para utilizar o método Backpropagation, foi necessário a normalização dos dados, visto que seus valores não podem ser muito discrepantes pois interferem diretamente no cálculo do reajuste dos pesos e piora a performance do método. Também auxilia na boa performance dos demais métodos.

O 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.

4 DESCRIÇÃO DA BASE DE DADOS

A base de dados utilizada na pesquisa contém informações clínicas de pacientes além de informações gerais como idade, sexo, etc. Por fim, essa base também apresenta o atributo target, que indica quais das instâncias possui ou não doenças cardíacas. Esses atributos são utilizados para a criação das regras de classificação utilizando os modelos citados anteriormente.

4.1 Seleção dos atributos

Os atributos selecionados foram escolhidos com base em estudos anteriores e relevância teórica para a previsão de doenças cardíacas. Cada atributo foi incluído no modelo devido à sua associação comprovada com o risco de doenças cardíacas ou seu papel protetivo. 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.
- **Sex:** Atributo binário, que indica o 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 gênero de doenças cardíacas.
- **CP - Dores no peito (1 - 4):** Atributo contínuo, que representa os tipos de dores no peito, ou seja, as mesmas estão relacionados com os sintomas de uma possível doença cardíaca.
- **trestbps (Pressão arterial em repouso):** Atributo discreto, cujo é um indicador do nível. 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.
- **chol (Colesterol sérico em mg/dl):** Atributo discreto, que representa um indicador de colesterol LDL, ou colesterol "ruim", está associado a um maior risco de problemas como infarto ou AVC quando elevado.
- **fbs (Açúcar no sangue em jejum):** Atributo binário, representa o estado de normalidade da glicemia em jejum.
- **restecg (Resultados eletrocardiográficos em repouso):** Atributo 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.
- **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).
- **exang (Angina induzida por exercício):** Atributo binário que representa uma dor na região do peito. O indivíduo com angina costuma sentir desconforto ou pressão abaixo do esterno.
- **oldpeak (Depressão do segmento ST induzida por exercício em relação ao repouso):** Atributo binário que representa a falta ou não de um segmento ST no ECG quando um indivíduo realiza atividade física.

- **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).

A Figura 1, representa, nessa base de dados utilizada no estudo, a relação de instâncias classificadas com Doenças Cardíacas ou não classificadas com Doenças Cardíacas. A tabela 1 representa a relação de instâncias utilizadas para treino e para testes e suas respectivas classes.

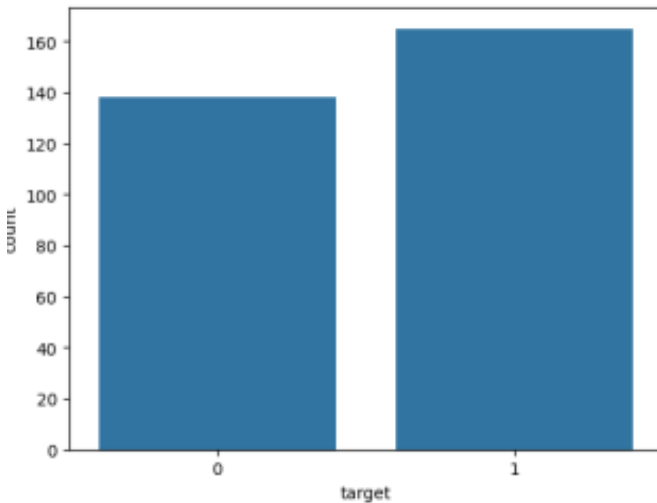


Figure 1: Relação de classes.

	Instâncias	Classe = 0	Classe = 1
Conjunto de treino	242	162	80
Conjunto de testes	61	38	23

Table 1: Distribuições Treino e Teste

5 METODOLOGIA

Para a modelagem, utilizou-se uma variedade de algoritmos de aprendizado de máquina, incluindo Decision Tree, KNN e Backpropagation, 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. Para estimar os melhores resultados, foram comparadas as métricas Acurácia, Precisão, Recall e F1-Score para avaliar o desempenho de cada modelo.

5.1 Avaliação de Desempenho

Métricas de avaliação de desempenho de modelos de aprendizagem de máquina são essenciais para definir qual o melhor modelo para um problema específico, indo além da análise apenas de acertos e erros. Para estimar os melhores resultados, foram comparadas as métricas Acurácia, Precisão, Recall e F1-Score para avaliar o desempenho de cada modelo.

- **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

$$\frac{VP}{VP + FP}$$

. 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

$$\frac{VP}{VP + FN}$$

. 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

$$\frac{2 \cdot \text{Preciso} \cdot \text{Recall}}{\text{Preciso} + \text{Recall}}$$

. 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.

5.2 Modelagem

Afim de definir o melhor modelo de aprendizado de máquinas entre os 3 escolhidos para o problema de classificação de doenças cardíacas, os mesmos foram aplicados na base de dados após o pré-processamento. Importante ressaltar que, para fins de melhorar os resultados, os mesmos conjuntos de treino e teste foram utilizados nos três modelos, assim como os mesmos processamentos.

Ademais, cada método foi escolhido por uma motivação bibliográfica. Todos os métodos realizam aprendizado supervisionado, que foi comprovado em outro artigo a respeito do mesmo tema como sendo o ideal para esse tipo de problema.

Os métodos Decision Tree e KNN foram testados no artigo ..., usado como referência nesse trabalho. Já o método Backpropagation foi escolhido baseado no estudo de redes neurais e com o objetivo de provar sua eficácia para esse tipo de problema.

- **Decision Tree:** As árvores de decisão são algoritmos de aprendizado supervisionado que podem ser usados tanto para classificação quanto para regressão. Sua interpretação é fácil e intuitiva porém estão propensas a overfitting.
- **KNN:** O KNN é um algoritmo de aprendizado baseado em instâncias que classifica novos casos com base na proximidade aos exemplos de treinamento no espaço de características. Pode ser computacionalmente caro para grandes conjuntos de dados, pois requer o cálculo de distâncias para cada previsão.
- **Backpropagation:** Backpropagation é um algoritmo utilizado para treinar redes neurais artificiais, ajustando os pesos das conexões com base no erro de previsão. No entanto, pode ser sensível à escolha de hiperparâmetros.

Para alcançar os resultados que serão discutidos na sequência, os modelos receberam hiperparâmetros baseados em referências bibliográficas de tópicos semelhantes.

Modelos	Hiperparâmetros
Decision Tree	criterion: entropy max_features: sqrt min_samples_leaf: 2 min_samples_split: 5
KNN	algorithm: auto metric: minkowski n_neighbors: 7 p: 2
Backpropagation	activation: relu batch_size: 32

Table 2: Hiperparâmetros Modelos de Aprendizado

Após a definição dos métodos a serem testados, configuração dos hiperparâmetros com base em pesquisas e referências a respeito do problema e pré-processamento da base de dados, foi realizado o treinamento dos modelos supervisionados com o intuito de classificar futuras instâncias. Com base nas métricas estabelecidas, é possível analisar os resultados, comparando os 3 métodos e entendendo qual é o ideal para a predição de doenças cardíacas.

6 RESULTADOS

Apesar da escolha dos algoritmos ser baseada em outros estudos e pesquisas, cada um deles possui questões que não os tornam ideais para o problema, e com base nos resultados dos modelos, é possível desenvolver essas questões.

Ao aplicar o método CART, foi possível observar um bom desempenho para o problema. O indicador "precisão" mostra que o modelo se saiu melhor classificando doenças cardíacas, já o indicador "recall" mostra que o modelo é capaz de

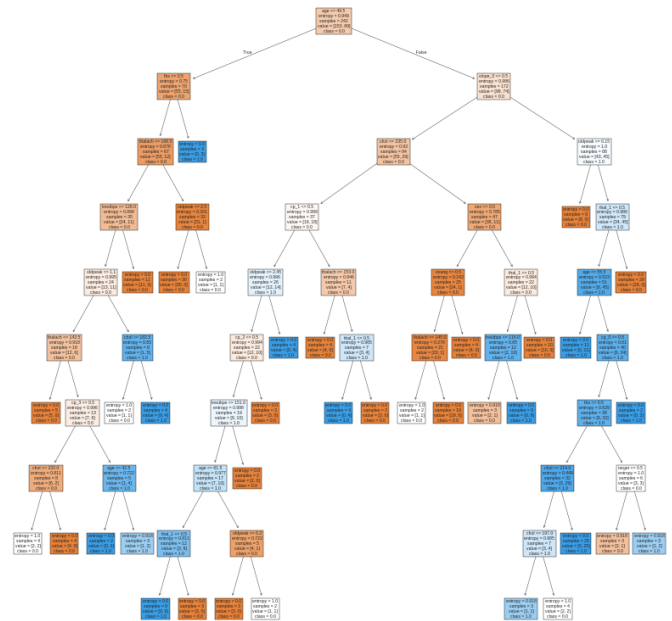


Figure 2: Decision Tree

	Classe = 0	Classe = 1
Precisão	0.86	0.92
Recall	0.94	0.82
F1 Score	0.90	0.87

Table 3: Métricas por Modelo - Decision Tree

capturar a maioria das instâncias não doentes, e a métrica "f1-score" se mostrou bem equilibrada para as duas classes.

	Classe = 0	Classe = 1
Precisão	0.94	0.87
Recall	0.88	0.93
F1 Score	0.91	0.90

Table 4: Métricas por Modelo - KNN

Já o método KNN apresenta um desempenho sólido, se adequando bem ao problema. O modelo apresenta maior eficácia em identificar as instâncias "não doentes", apresentando uma precisão maior para a classe 0, o "Recall" e o "F1-Score" também apontam para um ótimo desempenho do modelo, porém por se tratar de um problema no âmbito da saúde, espera-se uma maior eficácia na classificação dos indivíduos doentes.

O método Backpropagation mostrou-se consiso e eficaz, tendo em vista que todos os indicadores fora altos e bem próximos para ambas as classes, mostrando um equilíbrio necessário para um problema da área da saúde.

	Classe = 0	Classe = 1
Precis�o	0.88	0.89
Recall	0.91	0.86
F1 Score	0.90	0.87

Table 5: M tricas por Modelo - Backpropagation

O aprendizado de m quina   uma tecnologia de grande import ncia em diversas  reas na atualidade. A  rea da sa de se beneficia diretamente dos avanos da Intelig ncia Artificial como um todo, por m os m todos estudados nessa pesquisa sofrem de alguns problemas que devem ser observados com aten o ao aplicar-se no diagn stico de doenas.

O modelo de  rvore de decis o sofre com um problema de overfitting, que se trata de um ajuste exagerado aos dados de treinamento, sendo problem tico na hora de classificar novas inst ncias. Esse modelo tamb m   recomendado para bases menores de dados, e como a base estudada possui um n mero m dio de inst ncias e muitos atributos, a  rvore gerada ficou muito grande e de dif cil interpreta o.

J  o modelo KNN, apesar de ser ideal para conjuntos n o t o grandes de dados, j  que tem um gasto computacional alto, tamb m sofreu com outros problemas, como por exemplo a sensibilidade  s escalas dos dados. No caso desse estudo, o pr  processamento e os hiperpar metros foram definidos com base em outras pesquisas e trabalhos, o que possivelmente prejudicou o desempenho do m todo, mostrando que os modelos ainda precisam ser testados e estudados com mais intensidade na  rea da sa de.

O Backpropagation, m todo com menos informa es que os demais, tamb m foi impactado diretamente pela falta de informa es, j  que, por natureza, modelos de redes neurais s o intensivamente afetados pela escolha dos hiperpar metros.

6.1 Resultados e Compara es

	Acur�cia	M�dia	M�dia Ponderada
Decision Tree	0.89	0.88	0.88
KNN	0.90	0.90	0.90
Backpropagation	0.89	0.88	0.89

Table 6: Compara o entre os Modelos

Analisando os dados gerais de cada m todo   poss vel identificar uma melhor performance no m todo KNN. Isso ocorre devido ao fato de que os atributos das inst ncias tem valores pr ximos devido ao pr -processamento, facilitando e deixando o c lculo da dist ncia entre os pontos mais precisa. Al m disso, o KNN considera todos os atributos de forma igualit ria, diferente do CART, por exemplo, que pode ter

sofrido com o descarte de alguns atributos importantes ao longo do processo da constru o da  rvore.

Os resultados deste estudo s o relativamente consistentes com pesquisas anteriores que destacam a efic cia do KNN, sendo inferior apenas ao Random Forest, m todo que n o foi selecionado para o estudo por conta do objetivo de explorar modelos diferentes para classifica o de doenas. Al m disso, apesar do Decision Tree apresentar - se um pouco melhor em rela o ao KNN para diagn sticos positivos, observa - se neste estudo, que a suposi o de independ ncia condicional dos atributos pode limitar a efic cia deste algoritmo (CART) em conjuntos de dados com caracter sticas altamente conectados, como no campo da sa de.

7 CONCLUS O

A respeito da metodologia, esse trabalho foi capaz de reforar a import ncia de algumas fases da ci ncia de dados e estudo de Intelig ncia Artificial, como as etapas de pr -processamento, escolha correta dos hiperpar metros e a escolha de uma base de dados que se aplica ao problema ao mesmo tempo que   atual e relevante.

N o obstante, a vasta gama de trabalhos na  rea que foram utilizados como refer ncia para essa pesquisa demonstram que o uso de Aprendizado de M quinas para diagn stico de doenas   um t pico extremamente relevante na atualidade, focando na descoberta de modelos e configura es ideais para alcanar a maior precis o poss vel para diagnosticar doenas.

J  os m todos testados se mostraram extremamente eficazes para a tarefa proposta, sendo que a diferena entre os 3 foi m nima. Por m tamb m foi poss vel perceber que o meio da sa de   um desafio extremamente complexo para a Intelig ncia Artificial pois   esperado uma efic cia pr xima a 100, ja que tanto os valores classificados como Falso-Negativos quanto Falso-Positivos tem custo imensur vel por se tratar da vida das pessoas.

Portanto,   poss vel concluir que, apesar dos resultados promissores, existem in meras formas de aprimorar esse estudo, como a inclus o de novos atributos para compreender melhor esse fen meno, a cria o de novos algoritmos etc.

8 UTILIZA O DE IAS GENERATIVAS

Durante o desenvolvimento deste trabalho, a ferramenta GPT (Generative Pre-trained Transformer) foi utilizada de forma restrita para auxiliar algumas tarefas b sicas de formata o. Dessa forma, houve a utiliza o para verifica o de consist ncia textual e gera o de c digos para plotagem de gr ficos. Com isso, as  reas espec ficas onde o GPT foi empregado incluem: • Corre o ortogr fica e gramatical de trechos do texto. • Sugest es de formata o em LaTeX. • Aux lio na elabora o de se es introdut rias e conclusivas. • Gera o de c digos em Python.   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.

9 CÓDIGO DESENVOLVIDO

O código desenvolvido pode ser acessado nos seguintes links:

ACKNOWLEDGMENTS

Agradecimento especial aos colegas de grupo que estiveram envolvidos em todas as partes do projeto e a Professora Cristiane Neri, sempre compartilhando conhecimento teórico e prático dentro e fora de sala de aula.

REFERENCES