Inteligência Artificial: Estudo para Predição de Doenças Cardíacas

Análise e Aplicação no Conjunto de Dados 'Heart Disease Prediction' com Random Forest

Alexandre Augusto
PUC Minas
Belo Horizonte, MG

André Mendes
PUC Minas
Belo Horizonte, MG

Arthur Martinho
PUC Minas
Belo Horizonte, MG

Caio Gomes
PUC Minas
Belo Horizonte, MG

Daniel Salgado
PUC Minas
Belo Horizonte, MG

Rafael Maluf
PUC Minas
Belo Horizonte, MG

ABSTRACT

Este estudo investiga a modelagem preditiva de doenças cardíacas e tem como objetivo aprimorar a detecção precoce de doenças cardíacas por meio da análise dos atributos de uma base de dados, como idade, sexo, pressão arterial e níveis de colesterol. O artigo demonstra técnicas de pré-processamento que foram aplicadas para preparar os dados, além do algoritmo Random Forest, utilizado para classificação, alcançando uma acurácia de 87%. Embora o modelo identifique a maioria dos pacientes com doenças cardíacas, a presença de falsos negativos indica a necessidade de refinamento adicional.

CCS CONCEPTS

• **Machine Learning → Predictive Models**.

KEYWORDS

Doenças Cardíacas, Machine Learning, Pré-processamento de Dados, Random Forest

ACM Reference Format:
Alexandre Augusto, André Mendes, Arthur Martinho, Caio Gomes, Daniel Salgado, and Rafael Maluf. 2024. Inteligência Artificial: Estudo para Predição de Doenças Cardíacas: Análise e Aplicação no Conjunto de Dados 'Heart Disease Prediction' com Random Forest. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

O artigo se concentra na utilização do conjunto de dados "Heart Disease Prediction Dataset", disponível na plataforma Kaggle, para a construção de um modelo preditivo capaz de identificar a presença de doenças cardíacas em pacientes. A realização de uma análise detalhada desses dados melhora a compreensão dos fatores de risco associados às doenças cardíacas, e demonstra a eficácia de técnicas de pré-processamento, incluindo a limpeza de dados, a remoção de outliers, o balanceamento de classes e a redução de dimensionalidade, que são tópicos abordados ao decorrer do artigo.

2 ESTUDO DA BASE DE DADOS

A base de dados utilizada para realização do projeto foi a "Heart Disease Predicition Dataset", obtida através do site de base de dados "Kaggle" pelo senhor Krish Ujeniya. O conjunto de dados de predição de doenças cardíacas contém informações médicas utilizadas para prever a presença de doenças cardíacas em pacientes. Com um total de 303 registros e 14 atributos, esta base de dados inclui variáveis relevantes como idade, sexo, tipo de dor no peito, pressão arterial em repouso, níveis de colesterol, glicemia em jejum, resultados eletrocardiográficos em repouso, frequência cardíaca máxima alcançada, angina induzida por exercício, e depressão do segmento ST induzida por exercício em relação ao repouso. Os principais atributos contidos no conjunto de dados são:

- Idade: A idade dos pacientes varia de 29 a 77 anos.
- Sexo: A variável de sexo é binária, com 1 representando masculino e 0 feminino.
- Tipo de Dor no Peito (cp): Os tipos de dor no peito variam de 0 a 3.
- Pressão Arterial em Repouso (trestbps): A pressão arterial dos pacientes varia de 94 a 200 mm.
- Colesterol: Os níveis de colesterol variam entre 126 a 564 mg/dl.
- Glicemia em Jejum (fbs): A glicemia em jejum é superior a 120 mg/dl em 15% dos casos.
- Resultados Eletrocardiográficos (restecg): Esta variável varia de 0 a 2.
- Frequência Cardíaca Máxima (thalach): Os pacientes alcançam uma frequência cardíaca máxima que varia de 71 a 202 bpm.
- Angina Induzida por Exercício (exang): Aproximadamente 32,7% dos pacientes relataram angina induzida por exercício.
- Depressão ST (oldpeak): A depressão ST induzida por exercício varia de 0 a 6,2.
- Inclinação do Segmento ST (slope): A inclinação do segmento ST varia de 0 a 2.
- Número de Vessels (ca): O número de vasos sanguíneos coloridos varia de 0 a 4.
- Thalassemia (thal): Variando de 0 a 3.
- Alvo (target): Esta variável indica a presença (1) ou ausência (0) de doenças cardíacas, com uma média de 54,5% de pacientes apresentando a doença.

3 ETAPAS DE PRÉ-PROCESSAMENTO

O pré-processamento dos dados foi dividido em 4 seções para que todos os dados fossem treinados e testados de forma que o algoritmo estivesse apto para realizar os cálculos da melhor forma possível.

(1) **Limpeza de Dados**
Na primeira etapa do processo, é realizada a limpeza dos dados, fundamental para garantir a qualidade e a integridade das informações antes de qualquer análise. O conjunto de

Conference'17, July 2017, Washington, DC, USA
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

dados é carregado utilizando a biblioteca `pandas`, que facilita o uso de estruturas de dados. O método `dropna()` é utilizado para remover qualquer linha que contenha dados ausentes, evitando grandes distorções nos resultados das análises e dos modelos preditivos. Em seguida, `drop_duplicates()`- elimina valores duplicados no conjunto de dados, garantindo que cada registro seja único. A limpeza de dados foi escolhida como a primeira etapa do processo, pois dados imprecisos ou redundantes podem levar a conclusões erradas, reduzindo a eficácia dos modelos de *machine learning*.

(2) **Identificação e Remoção de Outliers usando IQR**
Na segunda etapa, abordamos a identificação e remoção de *outliers* (valores extremos que podem influenciar indevidamente a análise) usando o método do Intervalo Interquartil (IQR). O IQR é uma técnica estatística que ajuda a determinar a dispersão dos dados. A função `remove_outliers_iqr()` foi definida para calcular o primeiro quartil (Q1) e o terceiro quartil (Q3) de uma coluna específica, a partir dos quais calculamos o intervalo interquartil (IQR). Os limites inferior e superior são então definidos como $Q1 - 1.5 \times IQR$ e $Q3+1.5\times IQR$, respectivamente. A função remove os registros fora desses limites, o que ajuda a refinar o conjunto de dados. As colunas selecionadas para análise incluem `age` (idade), `trestbps` (pressão arterial em repouso), `chol` (colesterol), `thalach` (frequência cardíaca máxima) e `oldpeak` (depressão ST induzida por exercício). Após a remoção de *outliers*, o conjunto de dados tratado é salvo em um novo arquivo CSV.

(3) **Balanceamento de Dados com SMOTE**
A terceira etapa concentra-se no balanceamento do conjunto de dados utilizando a técnica de *oversampling* conhecida como SMOTE (*Synthetic Minority Over-sampling Technique*). Essa técnica é utilizada para lidar com a distribuição desbalanceada das classes no conjunto de dados, o que pode comprometer o desempenho dos modelos de predição. O conjunto de dados, após a limpeza e remoção de *outliers*, é dividido em variáveis independentes (X) e a variável dependente (y), que indica a presença ou ausência de doenças cardíacas. O conjunto de dados é dividido em subconjuntos de treino e teste usando a função `train_test_split()`. O SMOTE é então aplicado ao conjunto de treino para gerar exemplos sintéticos da classe minoritária, equilibrando assim as classes. Após a aplicação do SMOTE, a nova distribuição das classes é verificada e o conjunto de dados balanceado é salvo em um novo arquivo CSV.

(4) **Redução de Dimensionalidade usando PCA**
A última etapa aborda a redução de dimensionalidade utilizando a Análise de Componentes Principais (PCA), uma técnica que transforma um conjunto de variáveis correlacionadas em um conjunto de componentes principais não correlacionados. Esta técnica é útil em conjuntos de dados de alta dimensionalidade. O conjunto de dados balanceado é carregado e as variáveis independentes são normalizadas. O PCA é então aplicado para reduzir a dimensionalidade para duas dimensões. Um novo `DataFrame` é criado para armazenar os dados reduzidos, juntamente com a variável dependente. A visualização dos dados é feita utilizando um gráfico de dispersão, permitindo observar a separação das classes em relação aos dois componentes principais. A variância explicada por cada componente principal é calculada, fornecendo informações sobre a quantidade de informação retida após a redução de dimensionalidade.

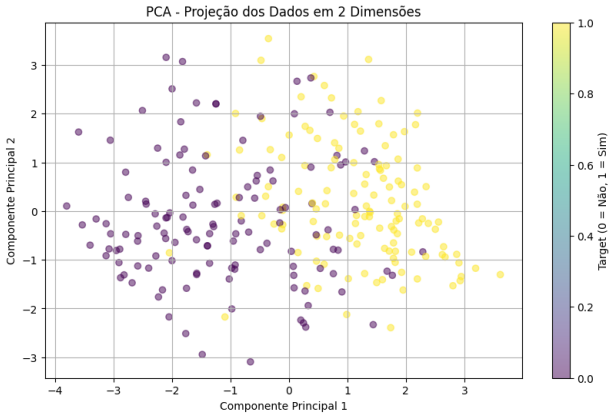

Figure 1: Projeção dos Dados com uso do PCA

4 ALGORITMO UTILIZADO

A Redução de Dimensionalidade (PCA) é uma técnica que reduz o número de variáveis, o que pode ajudar na visualização e simplificação do modelo. No entanto, pode resultar em perda de informação, o que não é sempre desejado antes de treinar modelos complexos como o Random Forest, com base nisso, colocamos o código do Random Forest antes do algoritmo de PCA.

O Treinamento do Modelo Random Forest é a quarta seção do algoritmo completo e foca no treinamento de um modelo de classificação utilizando o algoritmo Random Forest, por causa da sua robustez e capacidade de lidar com conjuntos de dados complexos. O arquivo CSV balanceado é carregado, e as variáveis independentes são separadas da variável dependente. O conjunto de dados é então dividido em subconjuntos de treino e teste. Para melhorar o desempenho do modelo, normalizamos os dados utilizando a classe StandardScaler, que transforma as características para que tenham uma média de 0 e um desvio padrão de 1, o que é especialmente importante para algoritmos baseados em distância. O modelo Random Forest é então instanciado e treinado com o conjunto de dados de treino. Após o treinamento, previsões são realizadas no conjunto de teste e o desempenho do modelo é avaliado usando métricas como acurácia, matriz de confusão e relatório de classificação. Esses resultados fornecem uma visão clara da eficácia do modelo em prever a presença de doenças cardíacas.

5 RESULTADOS

O algoritmo Random Forest produziu os resultados abaixo após utilizarmos um test-size = 0.35 e random-state = 42, ou seja, o test-size faz com que 65% das amostras sejam para treino e 35% para treino e o random-state faz com que toda vez que o código for executado, seja obtido a mesma divisão dos dados.

Table 1: Resultados do algoritmo

	Positivo	Negativo
Verdadeiro	36	4
Falso	42	8

A tabela abaixo é referente ao relatório de classificação gerado pelo algoritmo Random Forest. No algoritmo usamos 3 métodos: "accuracy = accuracy_score(y_test, y_pred)", "conf_matrix = confusion_matrix(y_test, y_pred)" e "class_report = classification_report(y_test, y_pred)".

accuracy = accuracy_score(y_test, y_pred) - tem como objetivo calcular a acurácia do modelo, utilizando da accuracy_scorem, que é uma função da biblioteca sklearn.metrics que mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões.

conf_matrix = confusion_matrix(y_test, y_pred) - tem como objetivo criar a matriz de confusão, utilizando confusion_matrix, que é outra função da biblioteca sklearn.metrics que fornece uma representação da performance do modelo de classificação, mostrando o número de previsões corretas e incorretas para cada classe.

class_report = classification_report(y_test, y_pred) - tem como objetivo gerar um relatório de classificação que resume as principais métricas de desempenho do modelo, utilizando classification_report, que é outra função de sklearn.metrics que gera um relatório que inclui precisão, recall, F1-score e suporte para cada classe.

Table 2: Relatório de Classificação

Classe	Precision	Recall	F1-score	Support
0	0.90	0.82	0.86	44
1	0.84	0.91	0.88	46
Accuracy	0.87			
Macro avg	0.87	0.87	0.87	90
Weighted avg	0.87	0.87	0.87	90

Os resultados obtidos com o algoritmo Random Forest, conforme detalhados na Tabela 1, indicam que o modelo conseguiu classificar corretamente 36 casos de doenças cardíacas (verdadeiros positivos) e 8 casos de pacientes saudáveis (verdadeiros negativos). No entanto, o modelo também apresentou 42 falsos negativos e 4 falsos positivos. Esses números ressaltam a importância de ajustar os parâmetros do modelo e de considerar diferentes abordagens para melhorar a precisão da classificação, especialmente em relação aos falsos negativos, que representam casos de pacientes com doenças cardíacas não detectadas pelo modelo. Esses resultados indicam que o modelo é mais eficaz em identificar corretamente pacientes com doenças cardíacas, embora ainda haja espaço para melhorias na identificação de pacientes saudáveis.

REFERENCES

[1] Arthur Martinho Caio Gomes Daniel Salgado Rafael Maluf Alexandre Augusto, André Mendes. 2024. Inteligência Artificial: Estudo para Predição de Doenças Cardíacas. Google Colab. https://colab.research.google.com/drive/1QrGxeEy7ktt1V504I8g8ILbET8DXPa8a?usp=sharing Acesso em: 29 set. 2024.

[2] Cristiane Neri Nobre. 2024. Ensembles Learning ou Aprendizagem em Conjunto. Slide presentes nos arquivos do Canvas. Acesso em: 29/09/2024.

[3] Cristiane Neri Nobre. 2024. ETAPAS DE PRÉ-PROCESSAMENTO BALANCEAMENTO DA BASE DE DADOS. Slide presentes nos arquivos do Canvas. Acesso em: 29/09/2024.

[4] Cristiane Neri Nobre. 2024. ETAPAS DE PRÉ-PROCESSAMENTO DADOS INCONSISTENTES E REDUNDANTES. Slide presentes nos arquivos do Canvas. Acesso em: 29/09/2024.

[5] Cristiane Neri Nobre. 2024. ETAPAS DE PRÉ-PROCESSAMENTO TRATAMENTO DE DADOS AUSENTES. Slide presentes nos arquivos do Canvas. Acesso em: 29/09/2024.

[6] Cristiane Neri Nobre. 2024. TRANSFORMAÇÃO DE DADOS REDUÇÃO DE DIMENSIONALIDADE. Slide presentes nos arquivos do Canvas. Acesso em: 29/09/2024.

[1] [2] [5] [6] [3] [4]