PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Machine Learning - Prof.: Cristiano Carvalho

TRABALHO ORIENTADO

Clinton Julio Novais Amaral

Flavia Azzi Nasser

Pedro Henrique de Souza Martins

Belo Horizonte

2021

a) A base de dados foi obtida do Kaggle (https://www.kaggle.com/fedesoriano/heart-failure-prediction) e trata-se de doenças cardiovasculares, que é a primeira causa de mortes global. Representa cerca de 31% de todas as causas mortis no mundo. Quatro de cinco mortes são de ataque cardíaco. Sendo que ⅓ destas mortes são prematuras e acomete pessoas abaixo dos 70 anos.

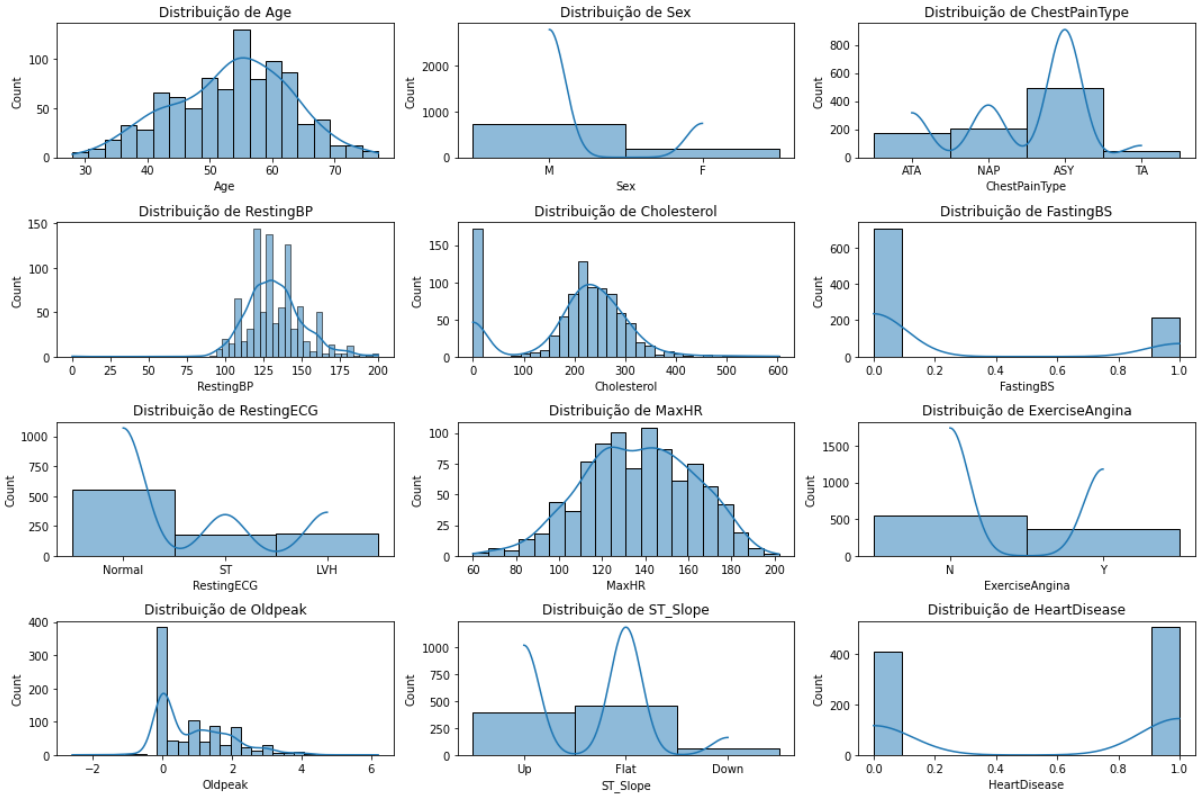
b) O modelo apresenta 11 características indicativas de causas de ataque cardíaco.

Uma vez que a doença se dá devido a fatores como diabetes, obesidade, pressão alta e outras pré -estabelecidas. Para tal a base de dados coletada, contém os seguintes índices as estas doenças relacionadas:

i) Idade: idade do paciente em anos
ii) Sexo: se M: masculino, F: Feminino
iii) Tipo de dor no peito: TA: Angina Típica, ATA: Angina Atípica, NAP: Sem dor de Angina, ASY: Assintomático]
iv) RestingBP: pressão arterial [mm Hg]
v) Colesterol: colesterol basal [mm/dl]
vi) FastingBS: indice glicemico [1: if FastingBS > 120 mg/dl, 0: otherwise]
vii) ECG em repouso: resultados de eletrocardiograma em repouso [Normal: Normal, ST: tendo anormalidade da onda ST-T (inversões da onda T e / ou elevação ou depressão de ST> 0,05 mV), HVE: mostrando provável ou definitiva hipertrofia ventricular esquerda pelos critérios de Estes]
viii) MaxHR: frequência cardíaca máxima alcançada [valor numérico entre 60 e 202]
ix) ExerciseAngina: angina induzida por exercício [S: Sim, N: Não]
x) Oldpeak: oldpeak = ST [valor numérico medido na depressão]
xi) ST_lope: a inclinação do pico do segmento ST do exercício [Up: uploping, Flat: flat, Down: downsloping]
xii) Doença cardíaca: classe de débito [1: doença cardíaca, 0: normal]

c) No ponto de vista do grupo, não haviam dados que poderiam ser descartados pois todos eram relevantes.

d) Base de dados está muito bem saneada, dados completos sem missing e bem definidos;



e) Fizemos a dumarização dos dados e repartimos 20% dos dados para utilizar como base de teste.

f) Utilizamos a árvore de decisão para a criação do nosso modelo. Utilizamos também randomSearch para verificar se era possível melhorar os hiperparâmetros.

g) Utilizamos a árvore de decisão pois a nossa base de dados possui a variável resposta qualitativa binomial que permite a utilização de um modelo classificatório. Após a criação do modelo inicial, verificamos que nosso modelo tinha overfitting. Com o intuito de aumentar a nossa acurácia, utilizamos o algoritmo randomSearch com 20 iterações para poder encontrar os melhores hiperparâmetros.

h) A utilização do randomSearch no proporcionou a criação de um modelo que conseguimos alcançar uma acurácia da base de treino de 0.87 e uma acurácia da base de teste de 0.82. O nosso melhor modelo criou uma árvore de três níveis de profundidade.

Tivemos um recall para (0) sem doença de 74% e (1) com doença de 88%.