I

Sistema de Predição de Doenças Cardíacas Utilizando o Classificador Ingênuo de Bayes

2021

ET586 | ESTATÍSTICA E PROBABILIDADE

II

NOSSO GRUPO

MARIA CLARA ACRUCHI

MARIA LUÍSA DOS SANTOS

VINÍCIUS SALES OLIVEIRA

Tópicos

- Introdução e Objetivos

Apresentação do tema abordado e os objetivos da nossa aplicação.

- Implementação e Métodos utilizados

Onde será mostrado como foi realizada a implementação e quais foram as técnicas e métodos utilizados para melhorá-la

- Experimentos e Testes

Abordaremos os experimentos executados, seus protocolos e quais tipos de validação foram usados.

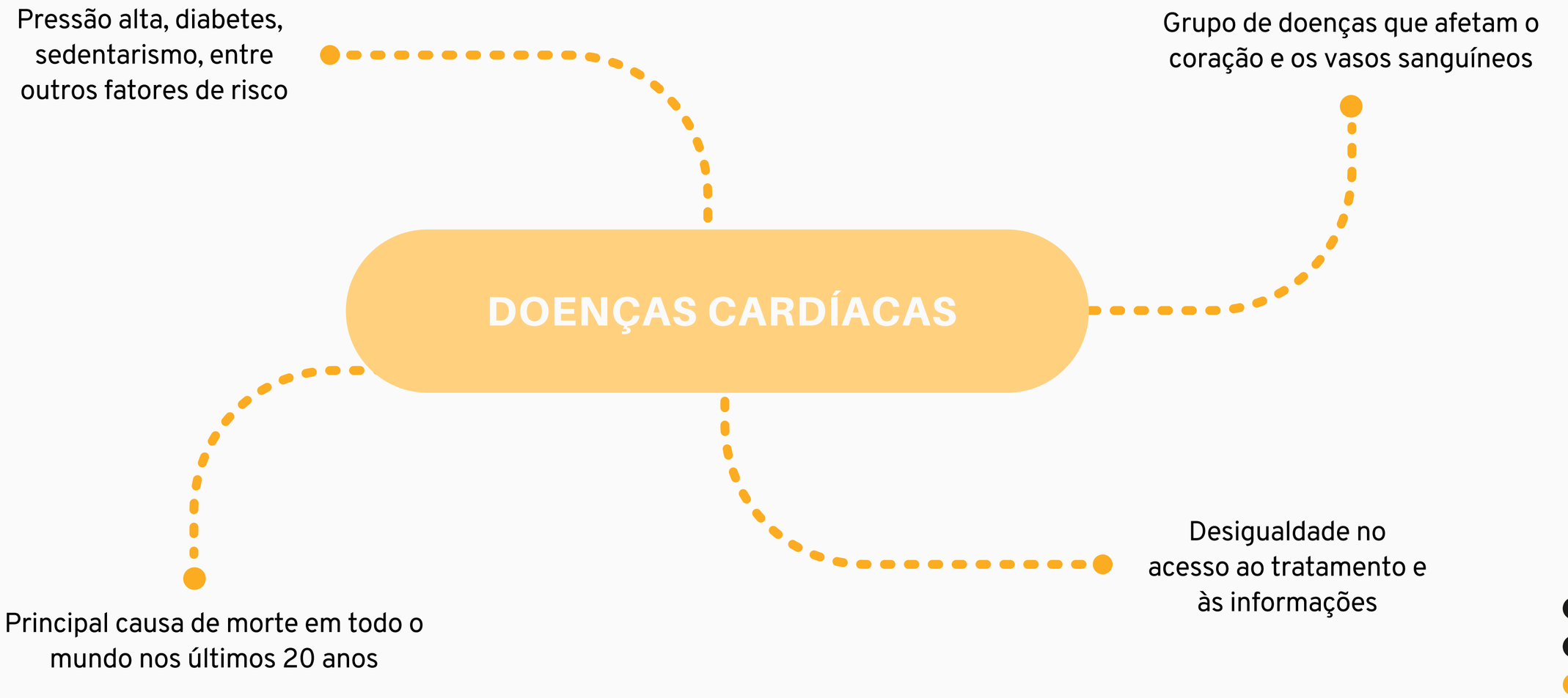
- Análise dos Resultados e Conclusões

Uma análise de validação da técnica será realizada para discutirmos as conclusões obtidas nesse processo.

IV

Introdução e Objetivos

Introdução


Pressão alta, diabetes, sedentarismo, entre outros fatores de risco
Grupo de doenças que afetam o coração e os vasos sanguíneos
DOENÇAS CARDÍACAS
Principal causa de morte em todo o mundo nos últimos 20 anos
Desigualdade no acesso ao tratamento e às informações

Objetivos

Predição de diagnóstico de cardiopatias

Construção de um sistema capaz de ajudar a reduzir o excesso de mortes, identificando características e padrões associados às doenças cardiovasculares em uma base de dados do repositório público UCI.

Data Science e Machine Learning

Utilização do Classificador Ingênuo de Bayes a partir de recursos e bibliotecas de aprendizagem de máquina para identificar esses padrões, no objetivo de inferir um diagnóstico de uma doença cardíaca.

Acesso menos desigual ao diagnóstico

Além de ser um utensílio que ajudará profissionais da saúde a tomar decisões clínicas mais rápidas e precisas do que os sistemas tradicionais de apoio podem oferecer.

VII

Implementação e Métodos utilizados

Base de Dados

DESCRIÇÃO DOS PARÂMETROS DA BASE DE DADOS

Atributo	Descrição
age	Idade em anos
sex	Valor 1: masculino. Valor 0: feminino
cp	Tipo da dor no peito. Valor 1: angina típica. Valor 2: angina atípica. Valor 3: dor não-anginosa. Valor 4: assintomático
trestbps	Pressão sanguínea em repouso medida em mmHg
chol	Colesterol sérico em mg/dl
fbs	Nível de açúcar no sangue em jejum >120mg/dl. Valor 1: verdadeiro. Valor 0: falso
restcg	Resultado de eletrocardiografia em repouso. Valor 0: normal. Valor 1: tem anormalidade ST-T. Valor 2: demonstra hipertrofia ventricular esquerda (LVH)
thalach	Frequência cardíaca máxima
exang	Angina induzida por exercício. Valor 1: sim. Valor 0: não
oldpeak	Depressão do segmento ST induzida por exercício em relação ao repouso
slope	Inclinação do oldpeak.Valor 0: ascendente. Valor 1: plano. Valor 2: descendente
ca	Número de vasos sanguíneos
thal	Determina o quão bem o sangue flui pela musculatura do coração. Valor 3: normal. Valor 6: fixed defect. Valor 7: reversable defect
num	Diagnóstico de doença cardíaca, é o estreitamento das artérias dado pelo resultado de uma angiografia. Valor 0: < 50% diameter narrowing. Valor 1: > 50% diameter narrowing

IX

Análise Exploratória dos Dados

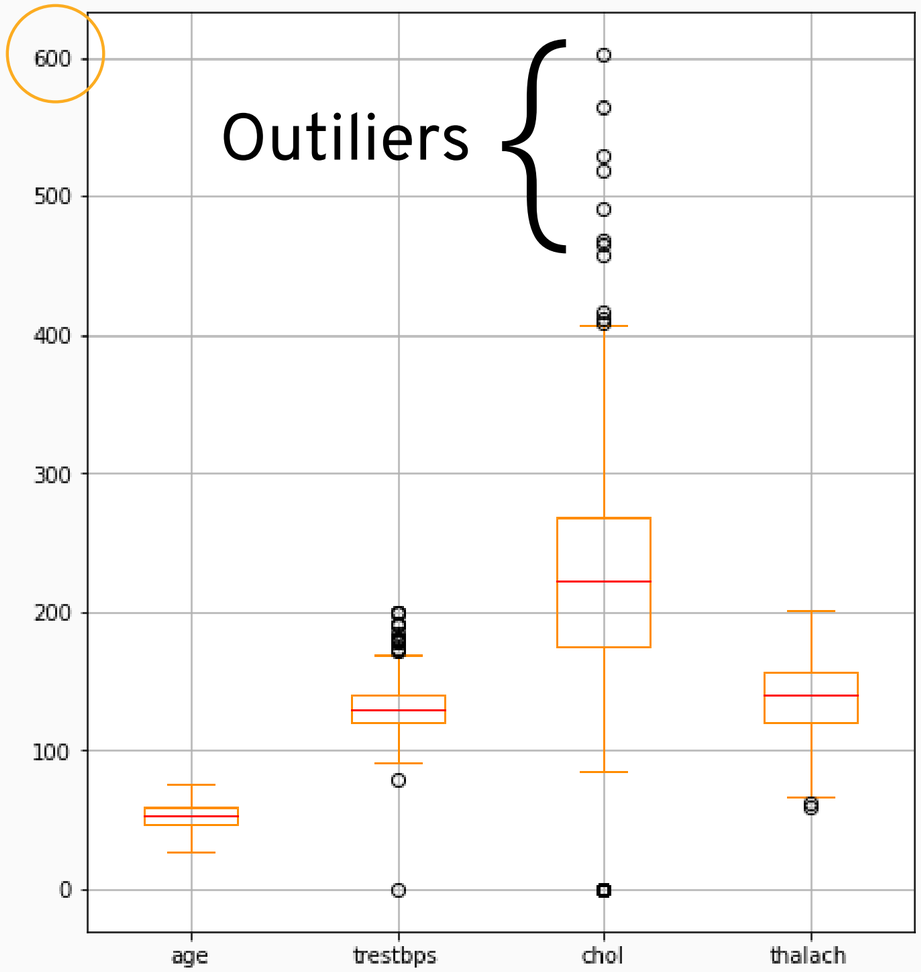
Entender o que cada variável representa

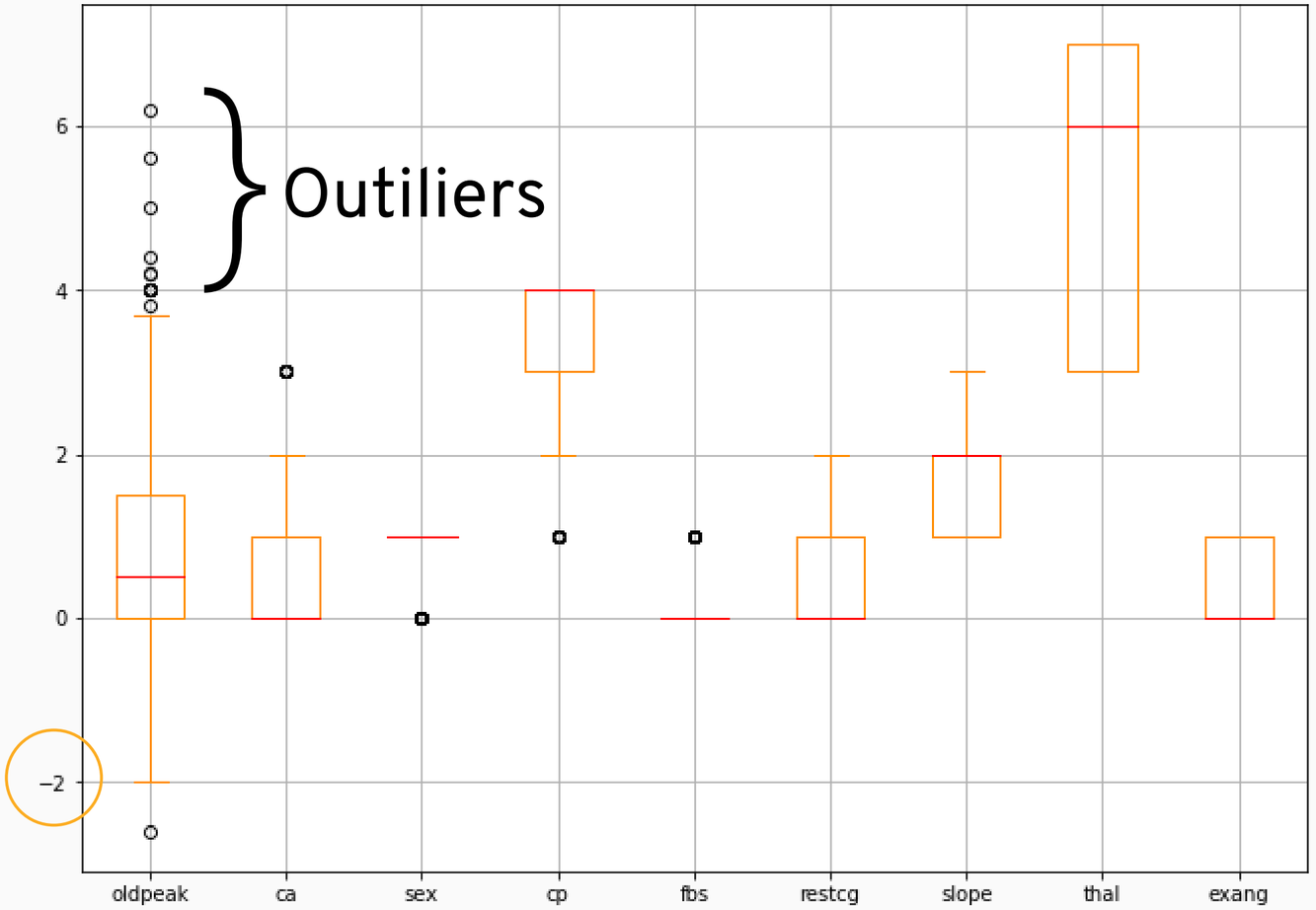
Nomear os parâmetros

Definir o tipo de cada variável

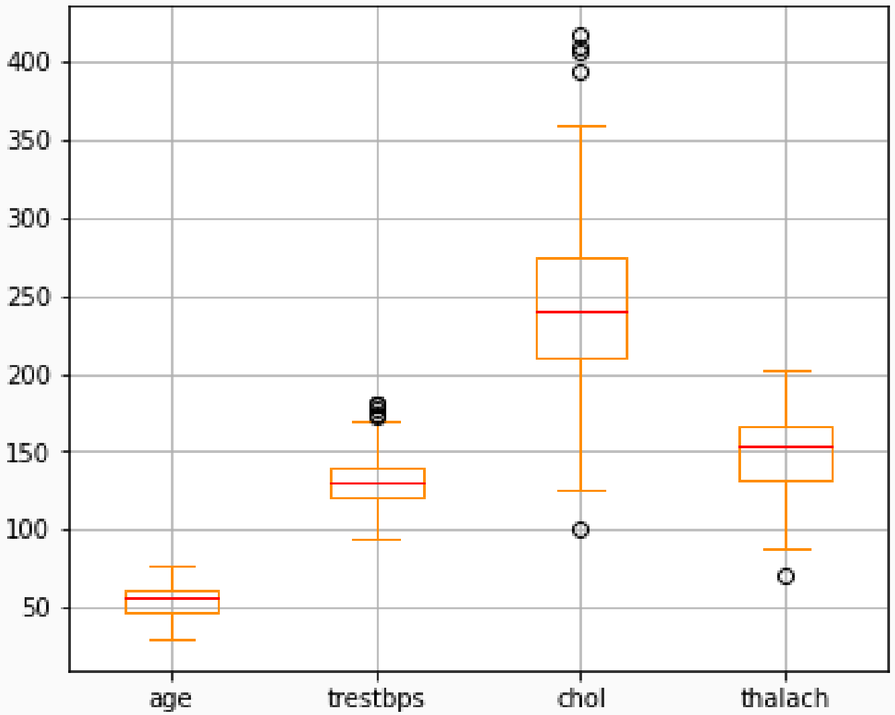
Tratar valores inválidos e espaços vazios

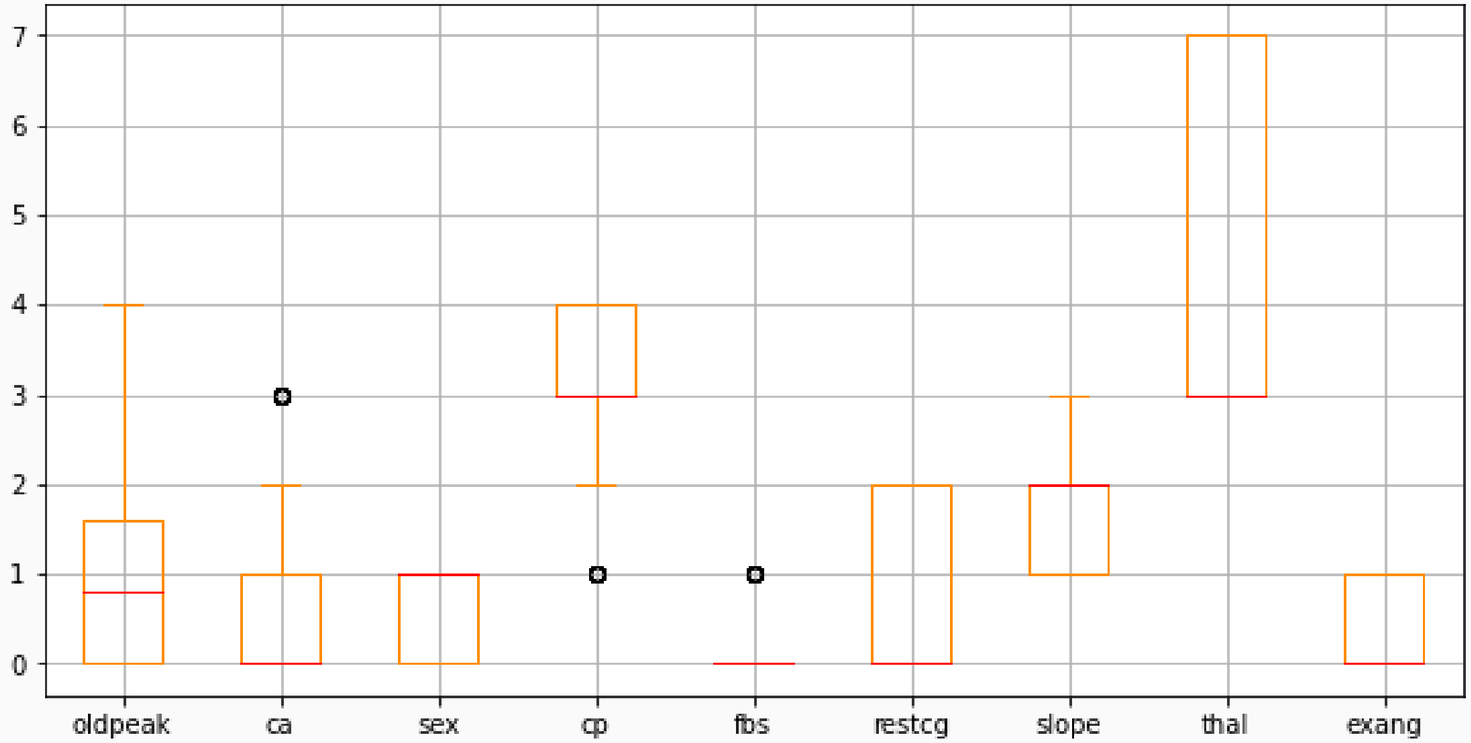
Remoção de Outliers





Remoção de Outliers





Remoção de valores inválidos

Valores de colesterol iguais a zero

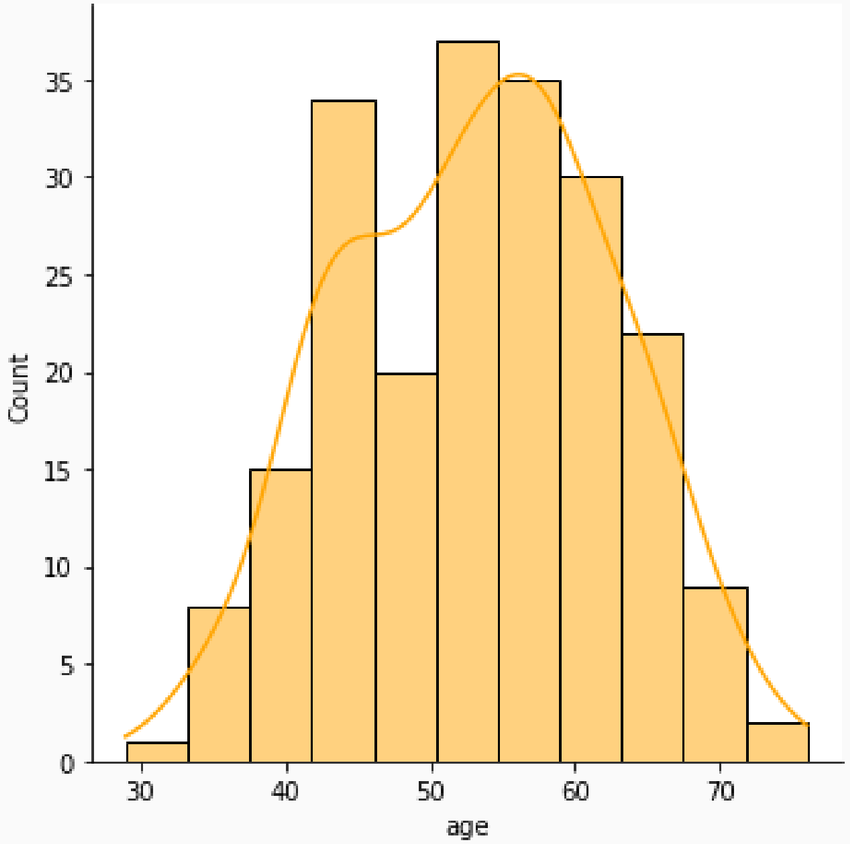
Substituição pela média dos valores da coluna

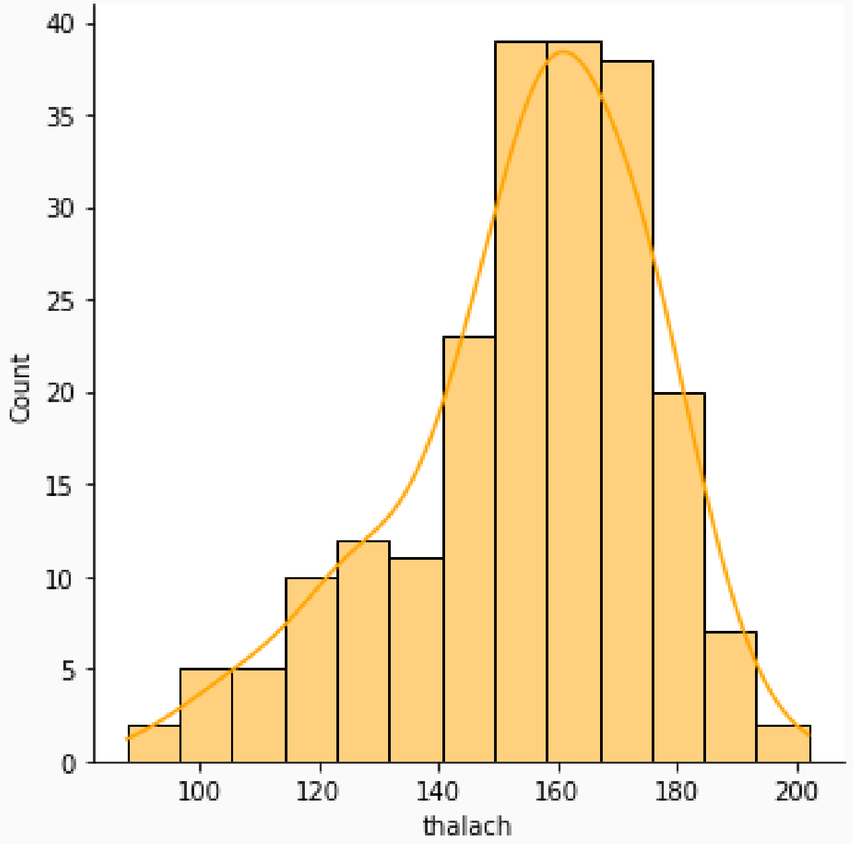
Valores não binários em uma coluna binária

As linhas com esses valores foram removidas

Gráficos e exemplos

Dados Numéricos

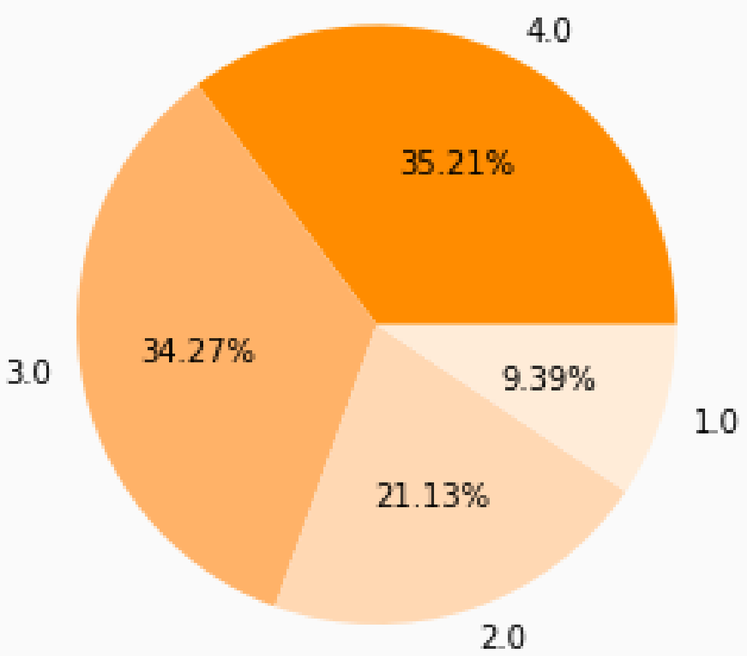




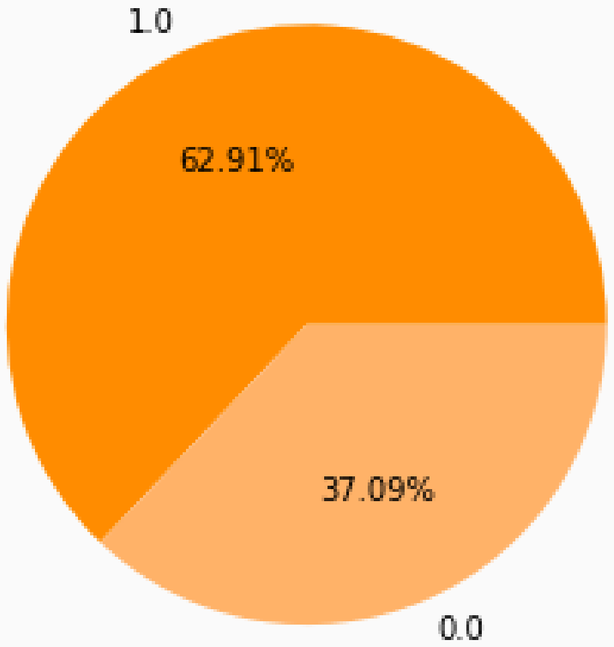
Gráficos e exemplos

Dados Categóricos

Tipo de dor no peito

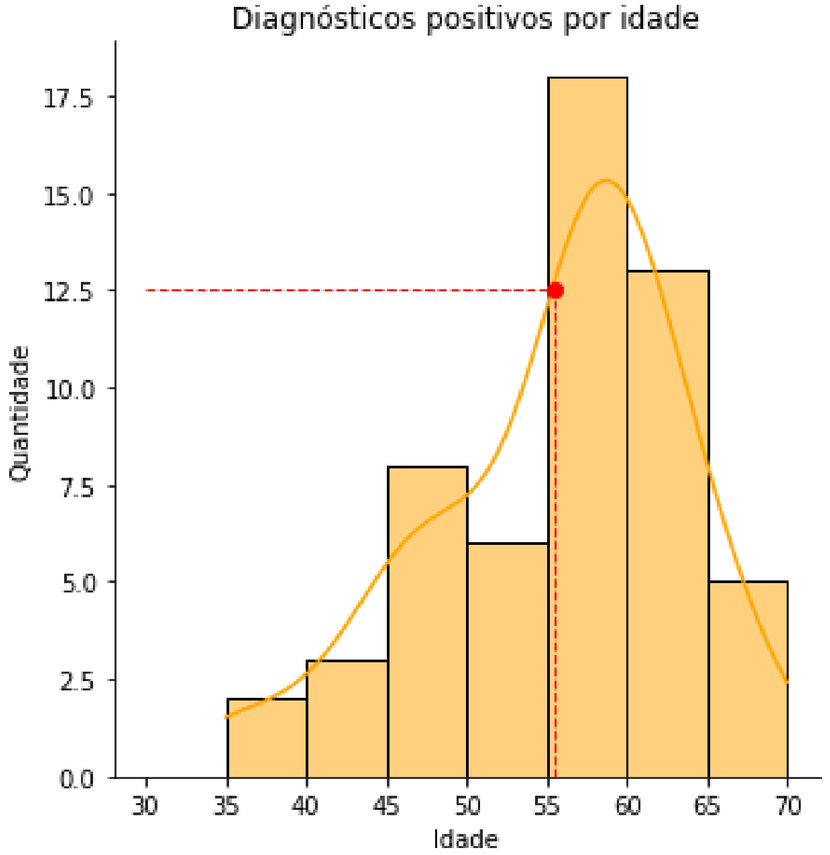


Sexo



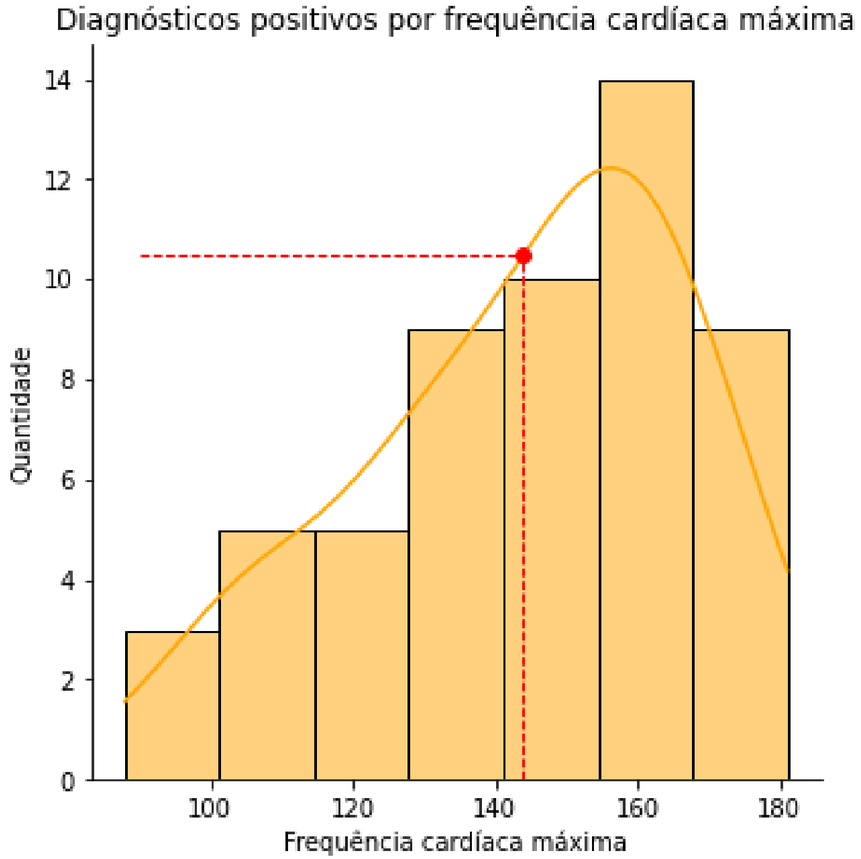
Análise Estatística

- Idade das pessoas diagnosticadas com doenças cardíacas:



Análise Estatística

- Frequência cardíaca máxima dentre os que têm diagnóstico positivo:



Classificador Ingênuo de Bayes

Principal conceito da Probabilidade Condicional

Método robusto e de alto desempenho

Fácil implementação e útil em grandes Data Sets

Supõe independência entre as variáveis

Identificação de padrões de ocorrência

XVIII

Classificador Ingênuo de Bayes

Aplicação com a
Distribuição Gaussiana

Pelos gráficos, os
atributos possuíam
Distribuição Normal

XIX

Análise dos Resultados

XX

Análise de Resultados

Validação Cruzada

Evita que o modelo seja adequado apenas para essa base de dados (overfitting)

Métricas e Indicadores

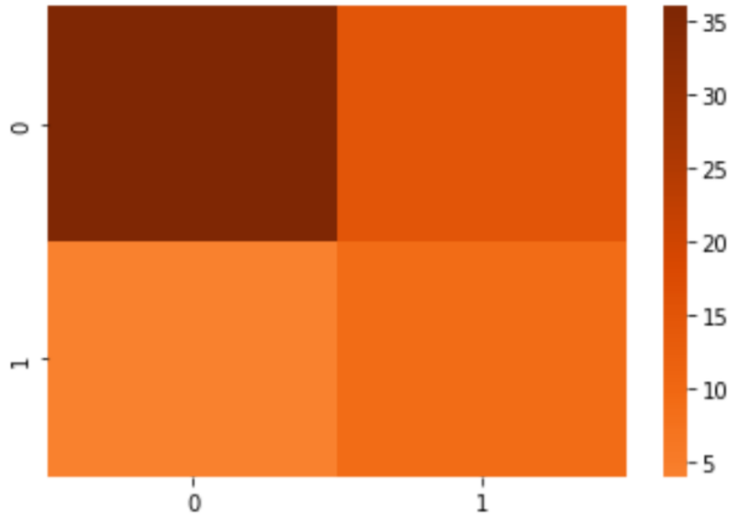
Recall

Precisão

Métricas e Indicadores

XXI

Matriz de Confusão



	Previsão de Diagnóstico Negativo	Previsão de Diagnóstico Positivo
Verdadeiro Negativo	36	15
Verdadeiro Positivo	4	9

XXIII

Experimentos e Análise dos Resultados

Métricas Analisadas

Recall é a métrica mais importante para a análise

Falsos negativos são muito mais prejudiciais no contexto de diagnóstico de doenças

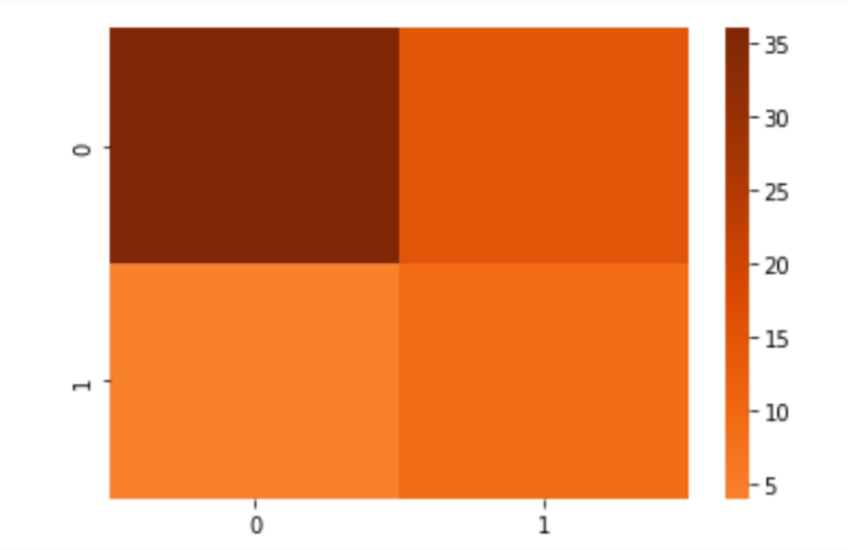
Experimento 1

É a base para os demais experimentos

Modelo treinado para toda a base de dados

Experimento 1

Resultados



	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.90	0.71	0.79	51
Positivo (1)	0.38	0.69	0.49	13
Acurácia	-	-	0.70	64
Média Macro	0.64	0.70	0.64	64
Média Ponderada	0.79	0.70	0.73	64

Experimento 2

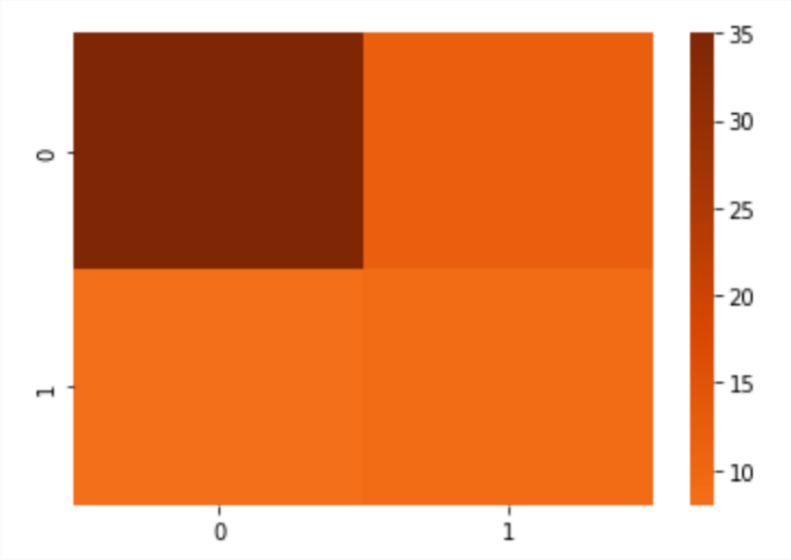
Estratificação dos pelos rótulos das classes

→

Observar o efeito de um conjunto de treinamento e de teste balanceados.

Experimento 2

Resultados



	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.81	0.74	0.78	47
Positivo (1)	0.43	0.53	0.47	17
Acurácia	-	-	0.69	64
Média Macro	0.62	0.64	0.63	64
Média Ponderada	0.71	0.69	0.70	64

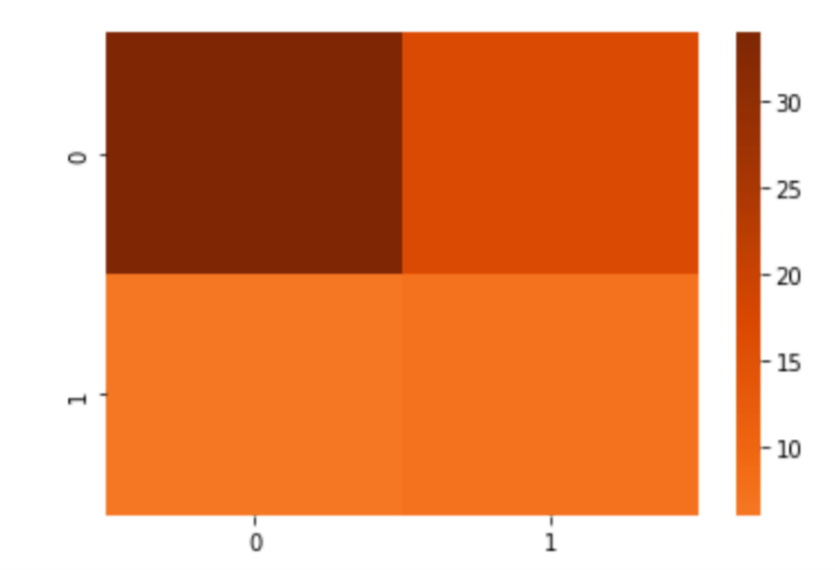
Experimento 3

Uso do algoritmo Boruta

Retorna um conjunto das melhores variáveis para se usar

Experimento 3

Resultados



	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.85	0.67	0.75	51
Positivo (1)	0.29	0.54	0.38	13
Acurácia	-	-	0.64	64
Média Macro	0.57	0.60	0.56	64
Média Ponderada	0.74	0.64	0.67	64

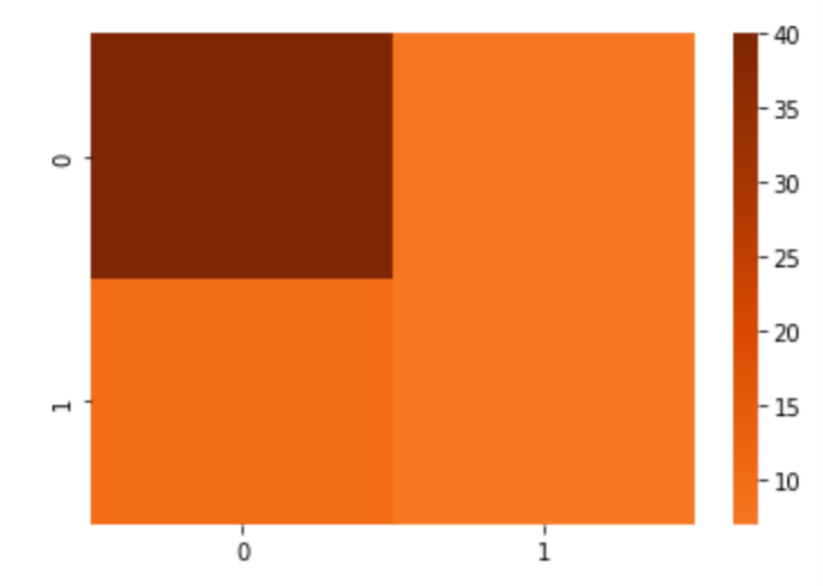
Experimento 4

Dados do experimento anterior foram estratificados antes do modelo ser treinado

Observar o efeito de um conjunto de treinamento e de teste balanceados.

Experimento 4

Resultados



	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.80	0.85	0.82	47
Positivo (1)	0.50	0.41	0.45	17
Acurácia	-	-	0.73	64
Média Macro	0.65	0.67	0.64	64
Média Ponderada	0.72	0.73	0.73	64

Conclusões e Discussões

Não prevê bem os diagnósticos positivos

Não atinge o objetivo inicial de ser um modelo capaz de prever a incidência de doenças cardíacas

Prevê bem os diagnósticos negativos

Embora não atinja o objetivo, o modelo é eficaz para o contexto mais arriscado: receber um diagnóstico negativo quando o indivíduo possui a doença, já que o número de falso negativos é mínimo