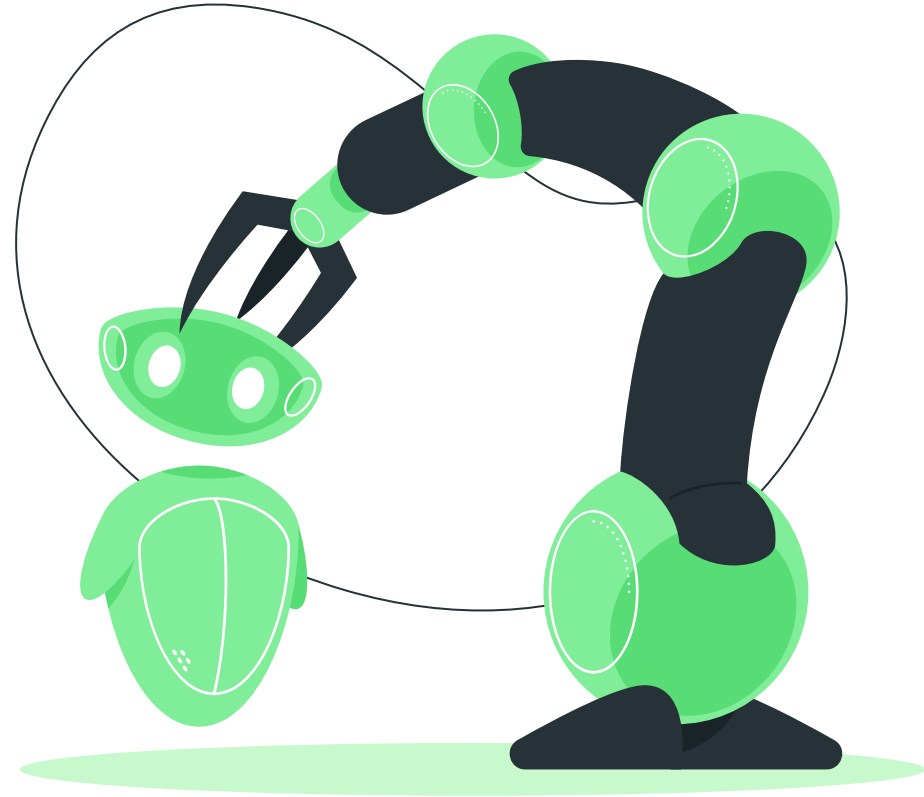


IA e Tópicos de BD

Daniele Maia Rodrigues – Professora

Gustavo de Oliveira Ventura 18562256
Ricardo Vieira 18238386

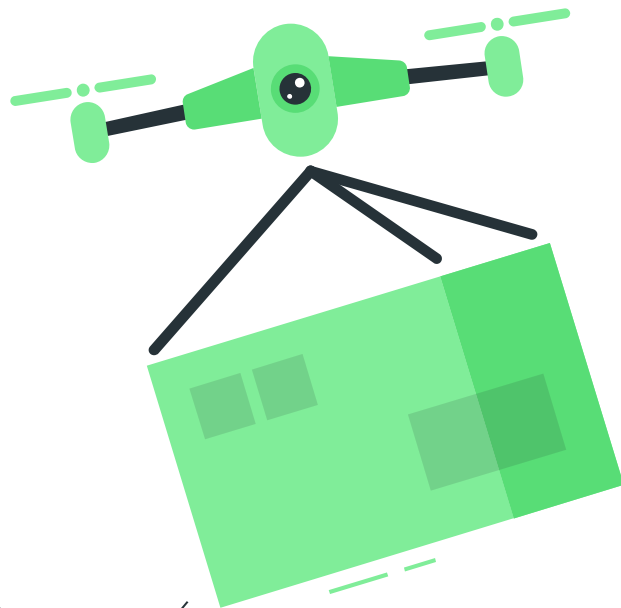


Introdução

Sobre o dataset:

Escolhemos o dataset “Cardiovascular Disease” que é relacionado a pacientes (Homens do Cabo Ocidental, África do sul) que tiveram ou não a doença coronária (infarto).

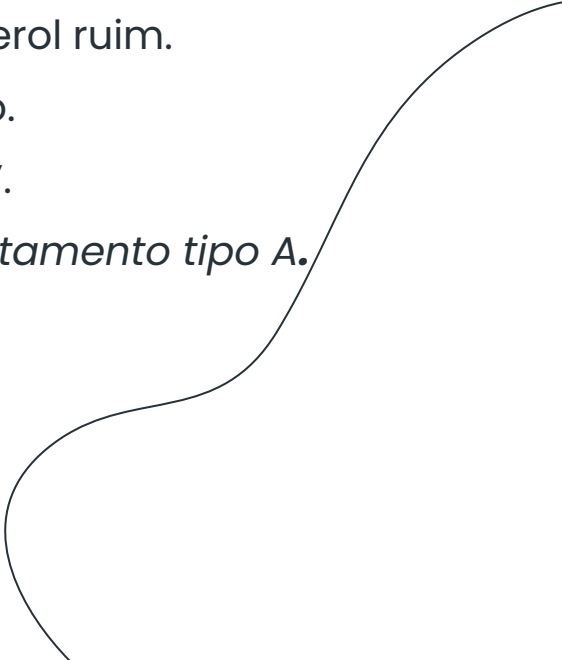
Esses dados são retirados de um conjunto de dados maior, descrito em Rousseauw et al, 1983, South African Jornal Médico.





Visualização de dados

Mais Sobre o dataset – Features

- **sbp (systolic blood pressure)** – Pressão arterial sistólica, pressão sanguínea em que o coração estimula bombeando no corpo.
 - **tobacco** - Nível de tabaco encontrado no sangue do paciente.
 - **ldl (low density lipoprotein cholesterol)** – Nível de colesterol ruim.
 - **adiposity** – Nível de gordura localizada no tecido adiposo.
 - **famhist** – Histórico familiar sendo “Presente” ou “Ausente”.
 - **typea** – Porcentagem de encaixe do paciente no *comportamento tipo A*.
 - **obesity** – Nível de obesidade do paciente.
 - **alcohol** – Nível atual de álcool no sangue do paciente.
 - **age** – Idade atual do paciente.
 - **chd** – Resposta para a doença coronária.
- 

Atividades iniciais

Buscando dados vazios

| | V1 | |
|-----------|----|--|
| ind | 0 | |
| sbp | 0 | |
| tobacco | 0 | |
| ldl | 0 | |
| adiposity | 0 | |
| famhist | 0 | |
| typea | 0 | |
| obesity | 0 | |
| alcohol | 0 | |
| age | 0 | |
| chd | 0 | |

Verificando os tipos de dados

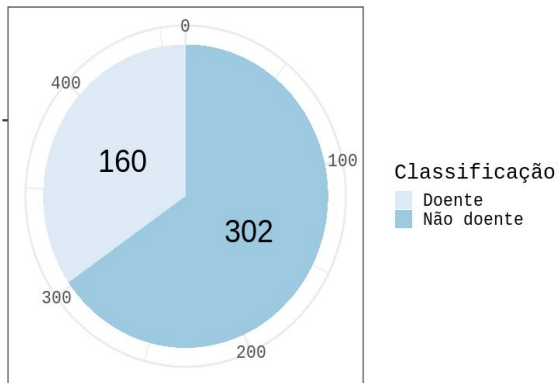
```
$ ind      : int  1 2 3 4 5 6 7 8 9 10 ...
$ sbp      : int 160 144 118 170 134 132 142 114 114 132 ...
$ tobacco  : num 12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
$ ldl      : num 5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
$ adiposity: num 23.1 28.6 32.3 38 27.8 ...
$ famhist  : chr "Present" "Absent" "Present" "Present" ...
$ typea    : int 49 55 52 51 60 62 59 62 49 69 ...
$ obesity  : num 25.3 28.9 29.1 32 26 ...
$ alcohol  : num 97.2 2.06 3.81 24.26 57.34 ...
$ age      : int 52 63 46 58 49 45 38 58 29 53 ...
$ chd      : int 1 1 0 1 1 0 0 1 0 1 ...
```

Dimensões

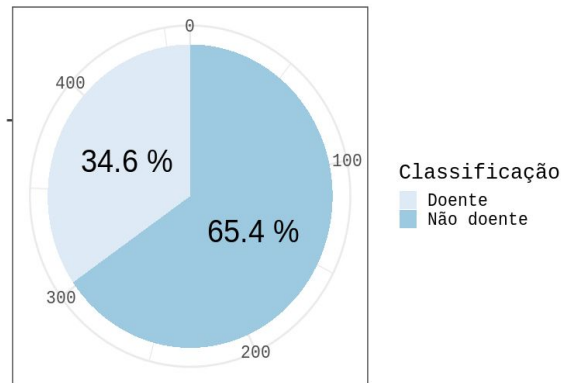
- 11 Features
- 462 Exemplares

Distribuição do dataset

Pie: Doentes e não doentes
Em dados brutos

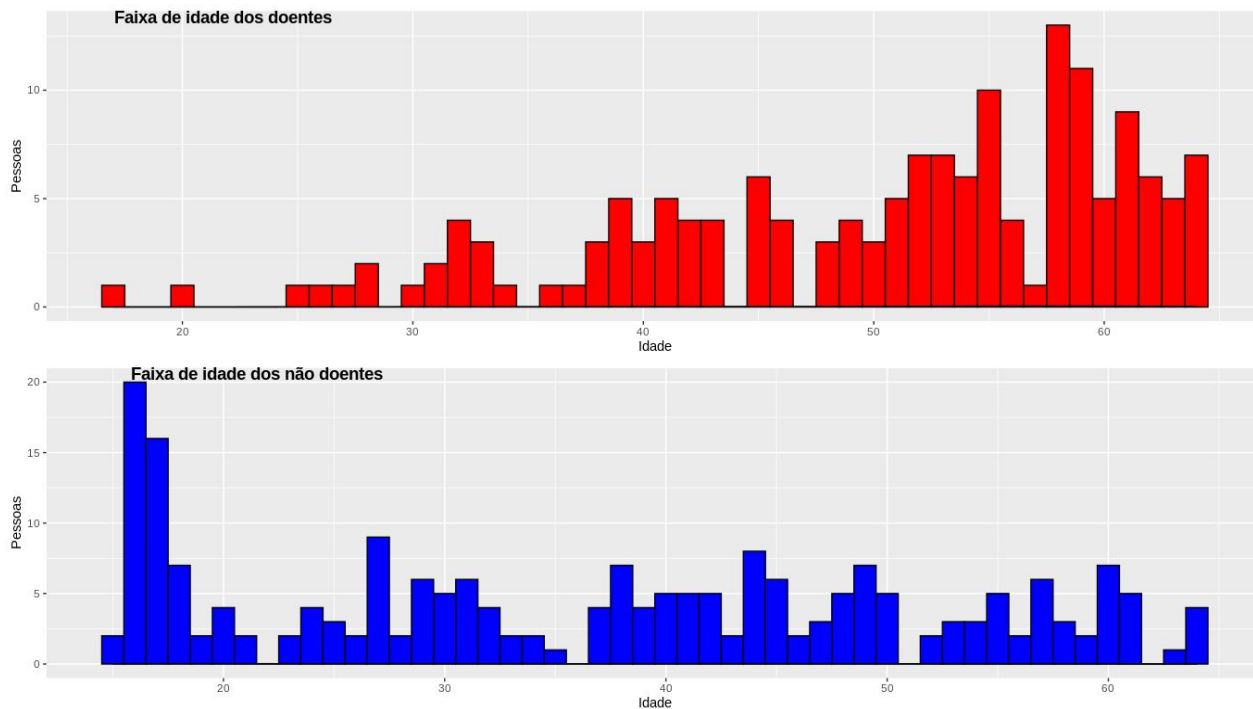


Pie: Doentes e não doentes
Em percentagem



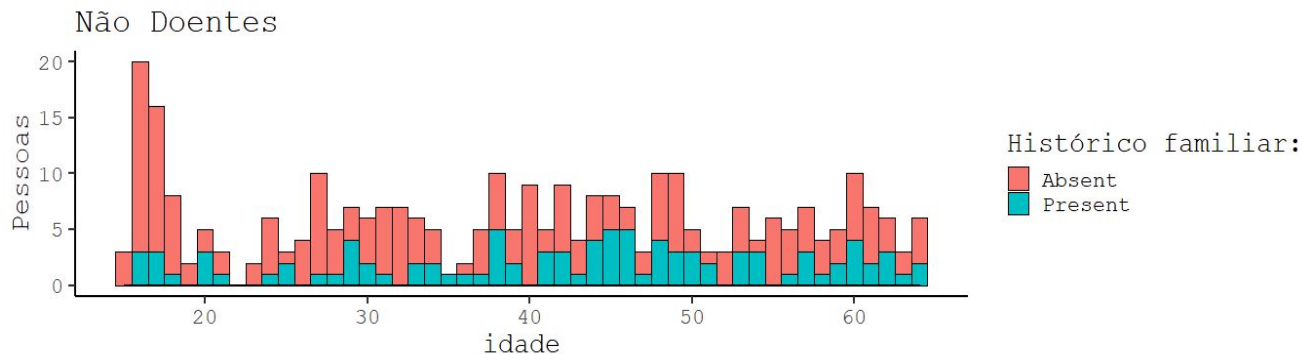
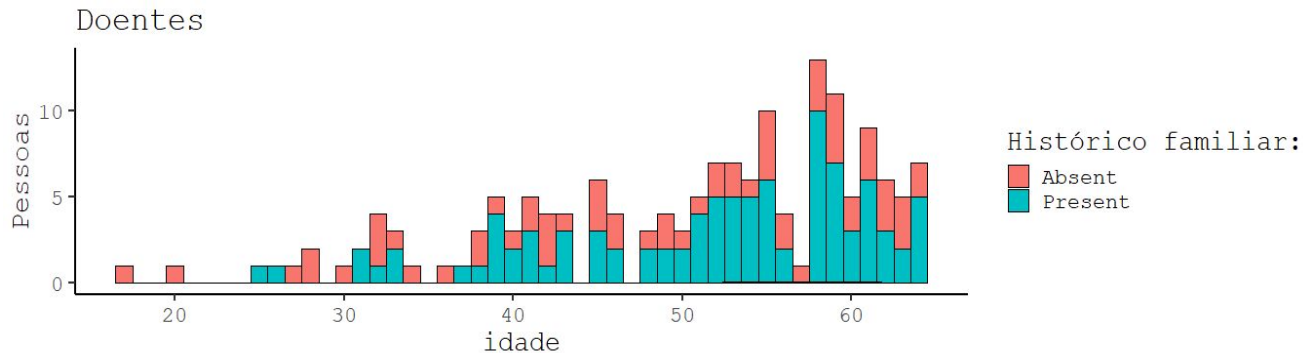
- Distribuição do dataset inicial
 - 160 Apresentando a doença;
 - 360 Não apresentado a doença.

Faixa de idade do Dataset



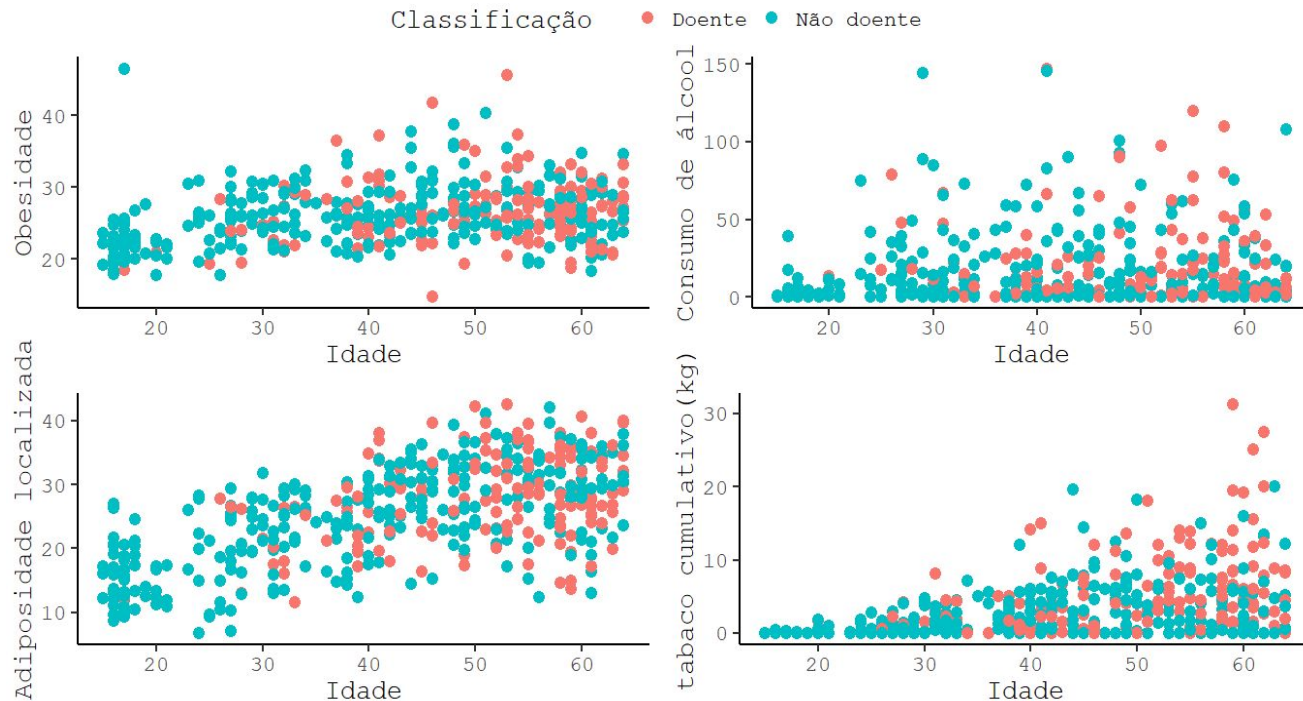
- Pico de ocorrências na faixa de idade próxima à 58, 59, 60 anos
- Adolescência com baixo índice de ocorrência na faixa etária entre 16 a 17 anos.

Faixa de idade do Dataset



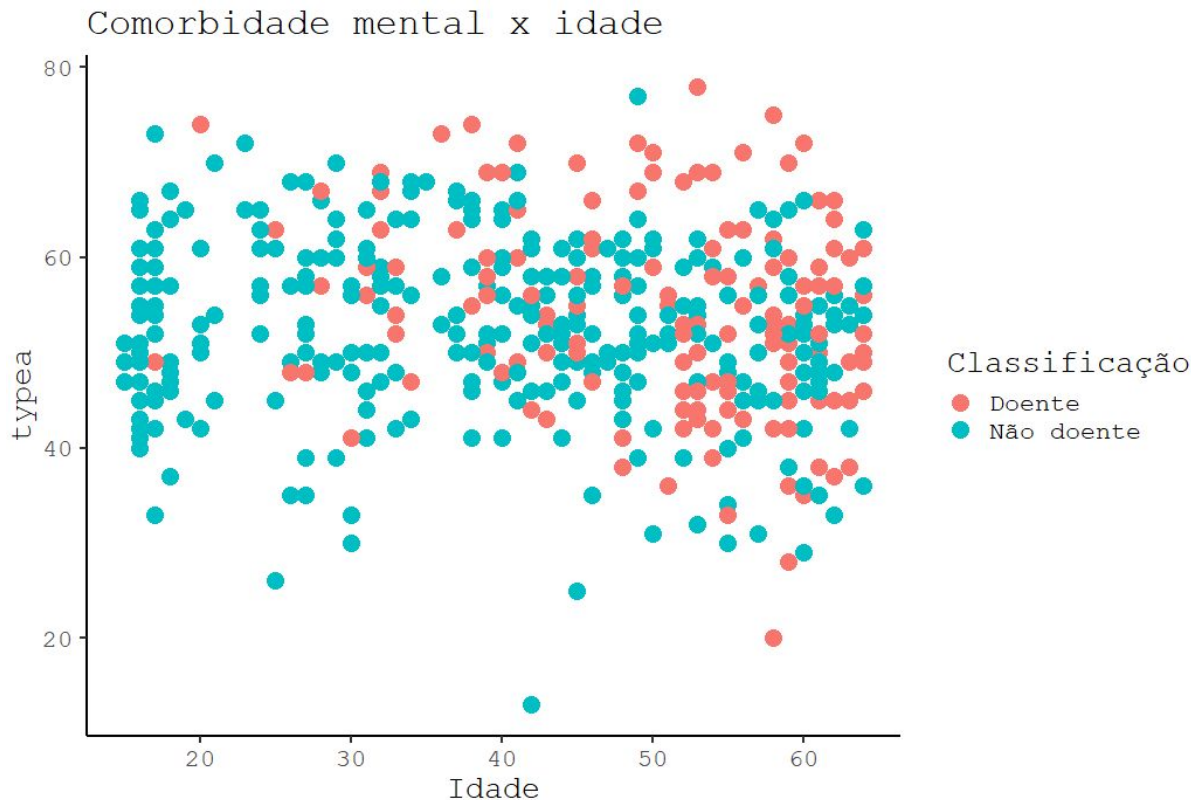
- Faixa de idade dos doentes e não doentes.
- Peso do histórico familiar.

Comorbidades físicas x idade



- Faixa mais perigosa/pré-disposta entre 40 - 60.
- Tabagismo e álcool na melhor idade podem desencadear a doença.

Comorbidade mental: Tipo A

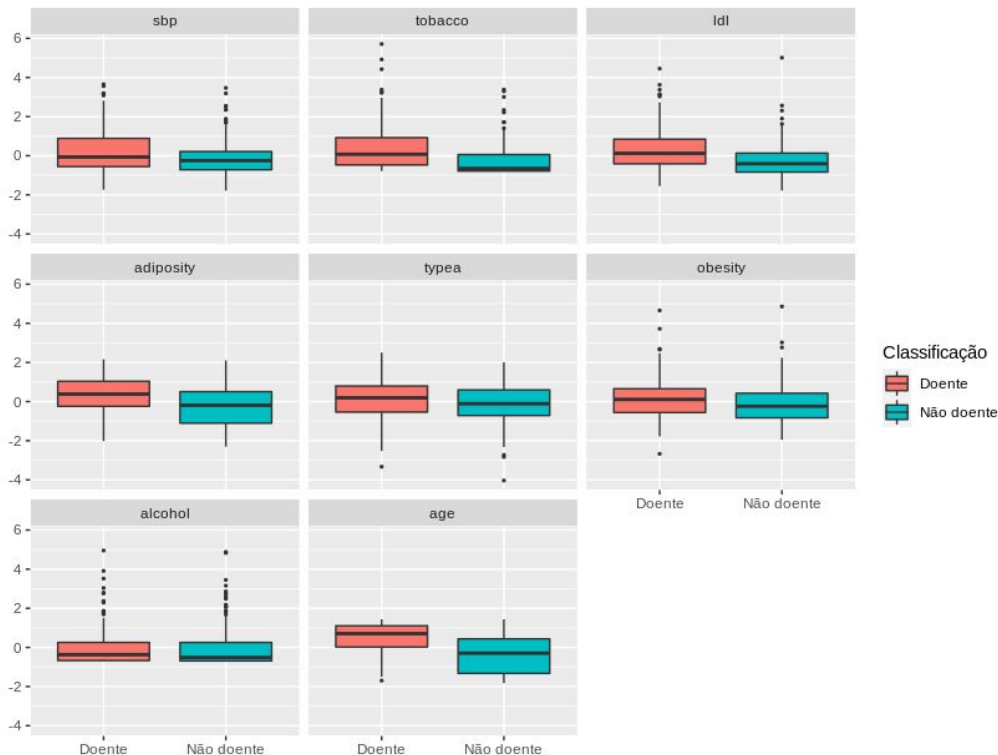


Características da personalidade comportamental A:

- Rigidamente organizado;
 - Impaciente;
 - Distúrbio de ansiedade;
 - Workaholic.
-
- Para os pacientes que se encaixam entre 40% e 60% nesse perfil, a partir dos 35 anos já é visível o aparecimento mais frequente da doença.

Comparativo de todas as features

Boxplot: Features normalizadas



- Fizemos uma normalização dos dados pois as unidades continham uma diferença muito grande, atrapalhando a visualização.
- Features *sbp*, *tobacco*, *ldl* e *adiposity* ganham destaque na diferença entre os pacientes doentes e não doentes;
- Visualização de possíveis candidatos a classificadores para a disciplina de Machine Learning.

Machine Learning



Atividade desenvolvida:

Classificação

Atividades desenvolvidas

Separação de conjunto de treino e teste

1

O dataset foi dividido em 80% treino e 20% teste para contemplar a aplicação dos algoritmos de classificação KNN e Árvore de decisão.

Aplicação do KNN

2

Aplicamos o algoritmo de classificação KNN em conjunto com os dataset de teste.

Árvore de decisão

3

Olhando para os nossos atributos, achamos aplicável ao algoritmo de árvore de decisão para gerar um classificador.

KNN – Resultados (1 e 3)

K sendo 1

| x | 0 | 1 |
|---|----|----|
| 0 | 42 | 16 |
| 1 | 22 | 13 |

Resultado: 59% de acurácia.

K sendo 3

| x | 0 | 1 |
|---|----|----|
| 0 | 42 | 16 |
| 1 | 26 | 9 |

Resultado: 54% de acurácia.

- Escolhemos como K, inicialmente números pequenos e ímpares, sendo eles: 1, 3, 5 e 11.

KNN – Resultados (5 e 11)

K sendo 5

| x | 0 | 1 |
|---|----|----|
| 0 | 46 | 12 |
| 1 | 27 | 8 |

Resultado: 58% de acurácia.

K sendo 11

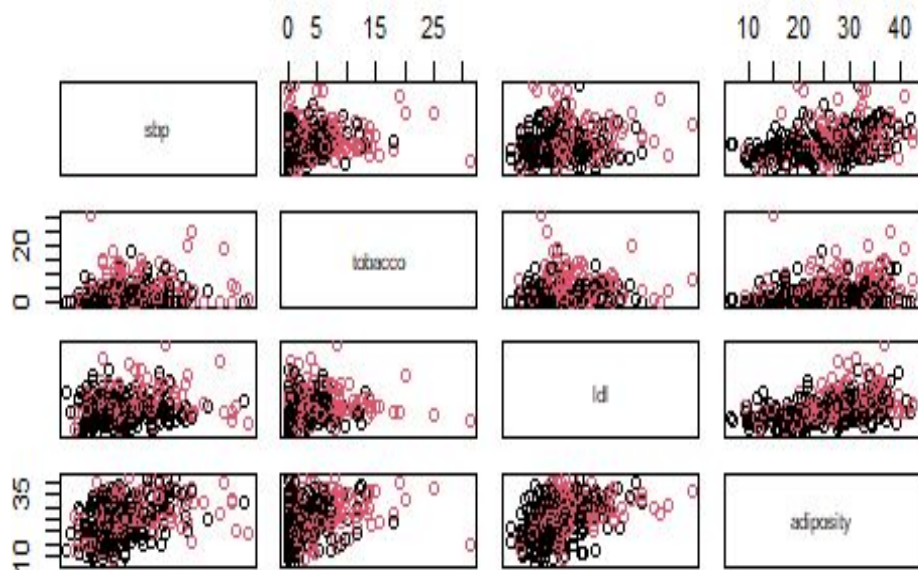
| x | 0 | 1 |
|---|----|---|
| 0 | 51 | 7 |
| 1 | 27 | 8 |

Resultado: 63% de acurácia.

- A partir de $k = 11$, os resultados começaram a decair em termos de acurácia.

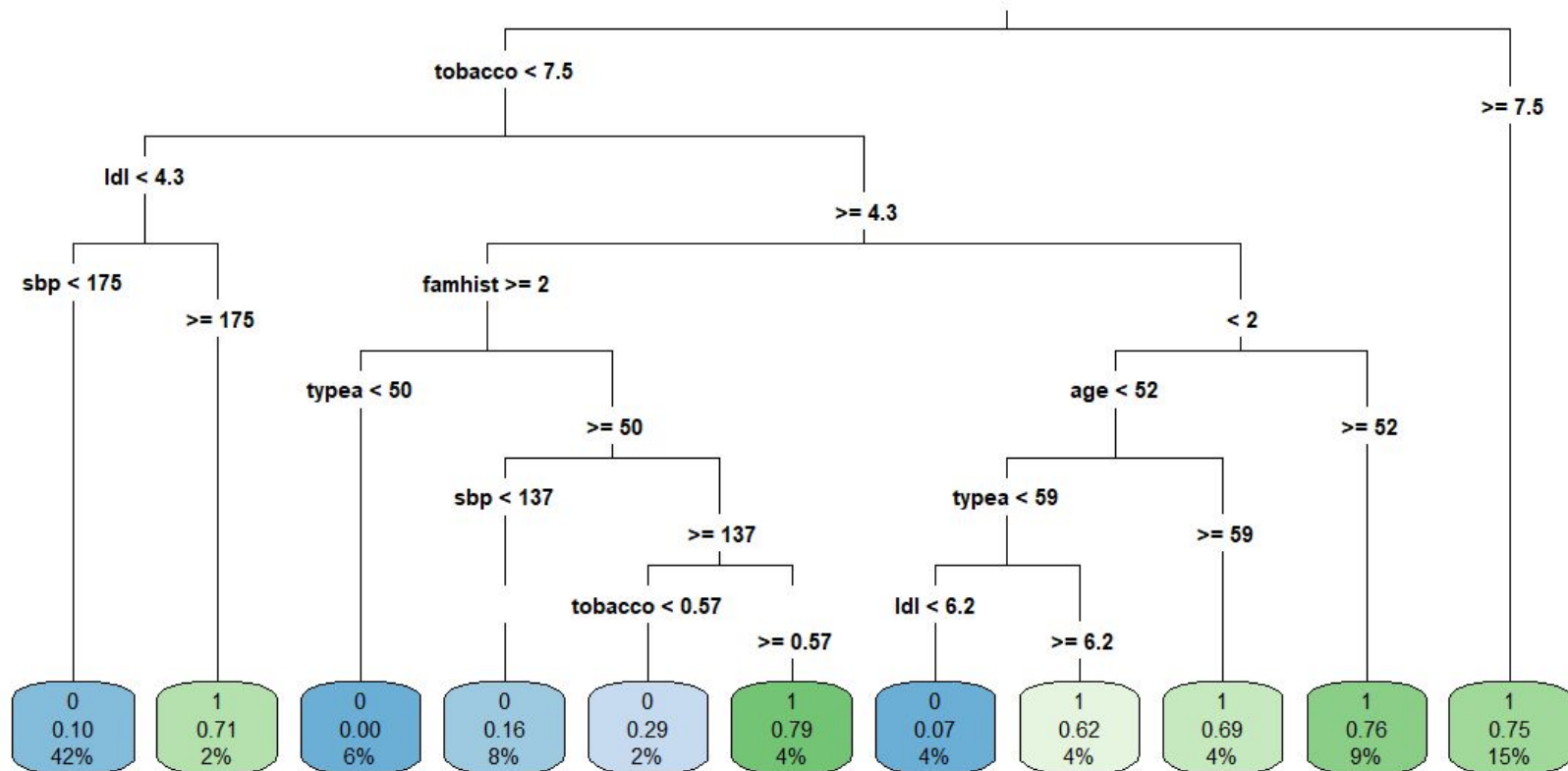
KNN – Análise

Minimizado:



- O motivo para que o algoritmo KNN tivesse a acurácia não tão alta como apresentado antes, seria a imprecisão do cálculo da distância euclidiana.
- Nenhuma feature apresenta uma separação clara das classes “Doente” e “Não doente”.

Árvore de decisão



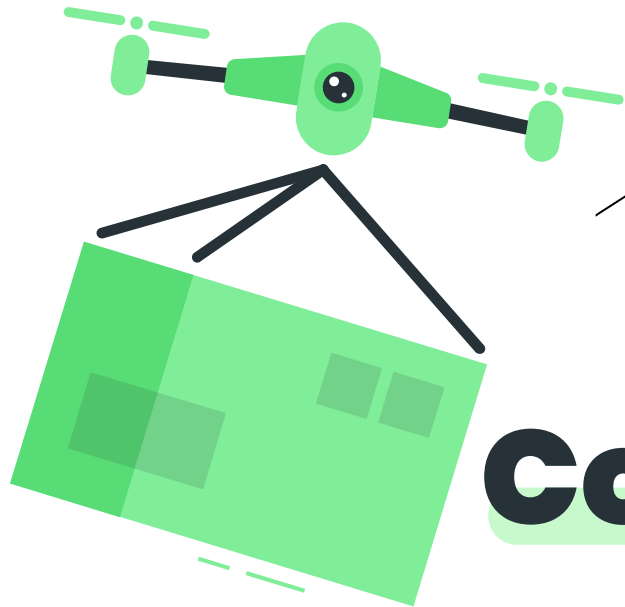
Árvore de decisão – Análise

Resultados e dataset

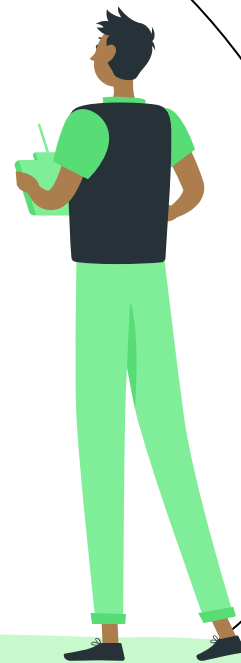
- Dataset de teste: 0.5913978 ou 59% índice de acerto
- Dataset de treino: 0.8346883 ou 83% índice de acerto
- Dataset de completo: 0.7857143 ou 78% índice de acerto

Features

- As melhores features que contribuíram para a pureza de cerca de 71% do classificador foram: O índice de **tabaco** no sangue e **ldl** (colesterol ruim).



Conclusão

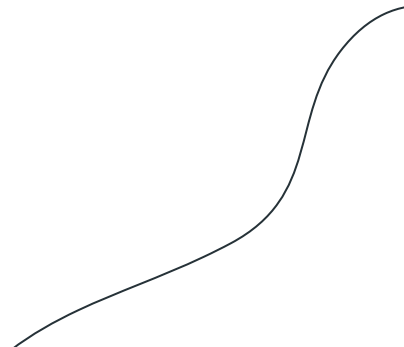


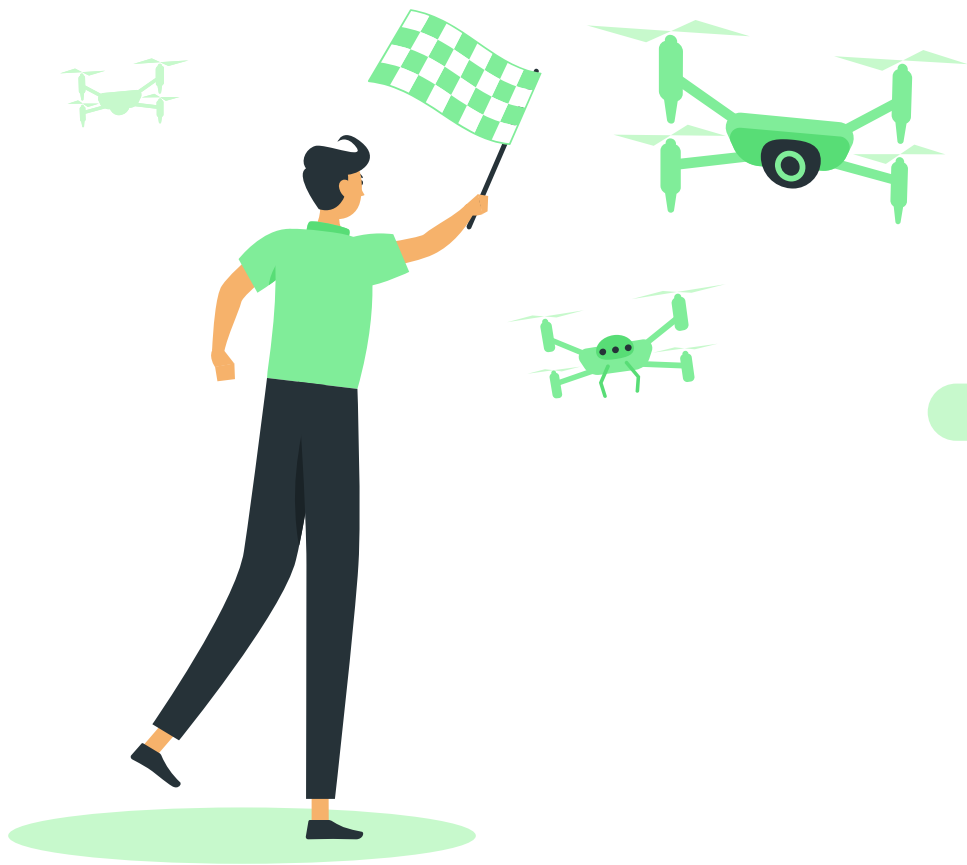
Conclusão e análise final

Machine Learning

- Comparando KNN e Árvore de decisão como classificadores obtivemos **59%** de acurácia para árvore de decisão e **63%** para KNN com $k = 11$, ambos analisados com o dataset de teste.

Visualização de dados

- Tendências relacionadas a faixa de idade e histórico familiar.
 - Comorbidades físicas e psicológicas são impactantes ao infarto.
- 



Obrigado!