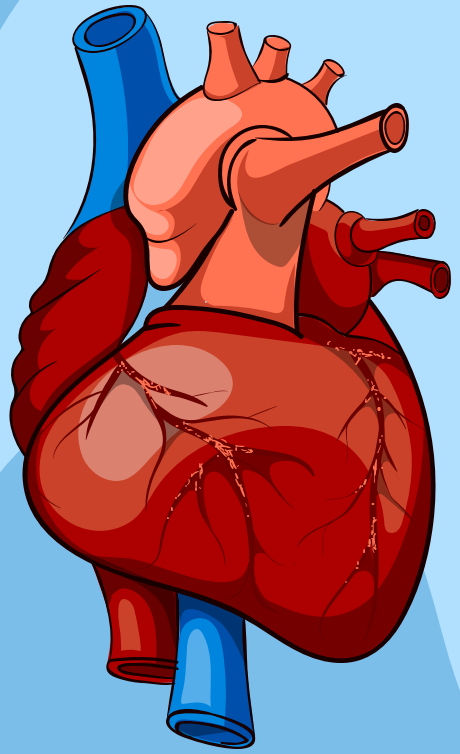


# Previsão de Insuficiência Cardíaca

Thiago Borsoni e Daniel Lloyd

# CONTEÚDO


01 **Introdução**

Apresentando o dataset

02 **Préprocessamento**

Análise exploratória

03 **Testes de normalidade**

Shapiro-Wilk

04 **Coeficiente de Correlação**

Pearson

05 **Regressão Linear**

Predição

06 **Regressão Logística**

Classificação

07 **Conclusão**

Conclusão do trabalho

01

# INTRODUÇÃO

O dataset foi escolhido por conter dados clínicos relevantes, como idade, tipo de dor no peito, colesterol e presença de doença cardíaca. Ele se destaca por atender aos requisitos do trabalho com variáveis contínuas e categóricas bem estruturadas. Além disso, aborda um tema de grande impacto social, permitindo a aplicação de machine learning em um contexto real e significativo.

01

**Número de registros:**
1190 observações

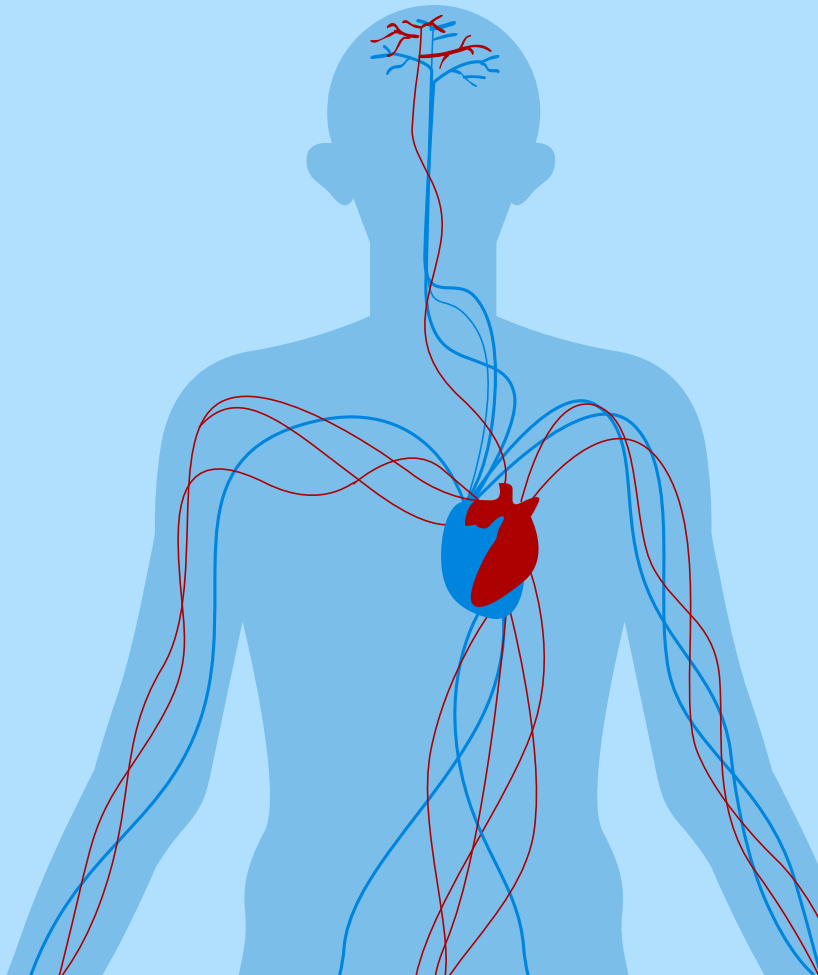
# Variáveis

- Age (Idade)
- Sex (Sexo)
- ChestPainType (Tipo de dor de peito)
- RestingBP (Pressão arterial em repouso)
- Cholesterol (Colesterol sérico)
- FastingBS (Glicemia em jejum)
- RestingECG (Resultados de eletrocardiograma em repouso)
- MaxHR (Frequência cardiáca máxima atingida [entre 60 e 202])
- ExerciseAngina (Angina induzida por exercício)
- Oldpeak (Medida de depressão do segmento ST)
- ST_Slope: (Inclinação do segmento ST no pico do exercício
- HeartDisease (Variável Alvo)

**Link:** https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

02

# Preprocessamento



# PREPROCESSAMENTO E ANÁLISE EXPLORATÓRIA

**0**

## Nulos

Nenhum dado nulo no dataset

**272**

## Duplicados

Removemos todos os dados duplicados

**0**

## Correlações Altas

Usamos praticamente todas as variáveis mas não precisamos remover nenhuma por correlação alta.

**0.40 / -0.40**

## Maior Correlação

Duas variáveis com maior correlação a target foram o MaxHR **(-0.40)** e Oldpeak (0.40)

# ESTATÍSTICAS

02

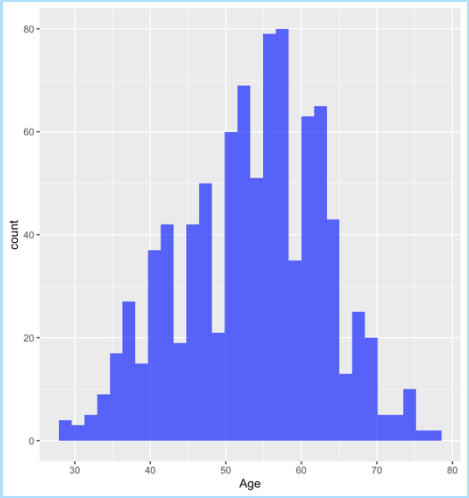
## Tipos

```
> str(heart)
'data.frame':   918 obs. of  12 variables:
 $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex           : chr  "M" "F" "M" "F" ...
 $ ChestPainType : Factor w/ 4 levels "ASY","ATA","NAP",..: 2 3 2 1 3 3 2 2 1 2 .
 $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
 $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
 $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
 $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
 $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

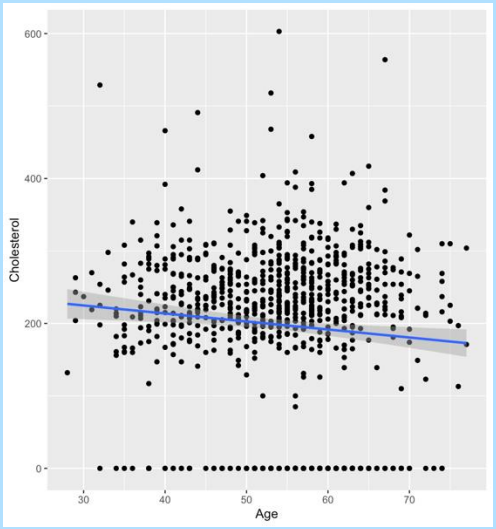
## Medidas Estatísticas

```
> summary(heart)
      Age            Sex            ChestPainType   RestingBP      Cholesterol      FastingBS       RestingECG            MaxHR
 Min.   :28.00   Length:918         ASY:496       Min.   :  0.0   Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
 1st Qu.:47.00   Class :character   ATA:173       1st Qu.:120.0   1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   NAP:203       Median :130.0   Median :223.0   Median :0.0000   Mode  :character   Median :138.0
 Mean   :53.51                      TA : 46       Mean   :132.4   Mean   :198.8   Mean   :0.2331                      Mean   :136.8
 3rd Qu.:60.00                                    3rd Qu.:140.0   3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
 Max.   :77.00                                    Max.   :200.0   Max.   :603.0   Max.   :1.0000                      Max.   :202.0
 ExerciseAngina       Oldpeak          ST_Slope          HeartDisease
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
                    Mean   : 0.8874                      Mean   :0.5534
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000
                    Max.   : 6.2000                      Max.   :1.0000
```

02




A distribuição de idade é **assimétrica à esquerda (levemente)**, com maioria dos pacientes entre **45 e 65 anos**. Essa faixa etária é coerente com o grupo de risco de doenças cardíacas, o que **torna o dataset bem contextualizado**.

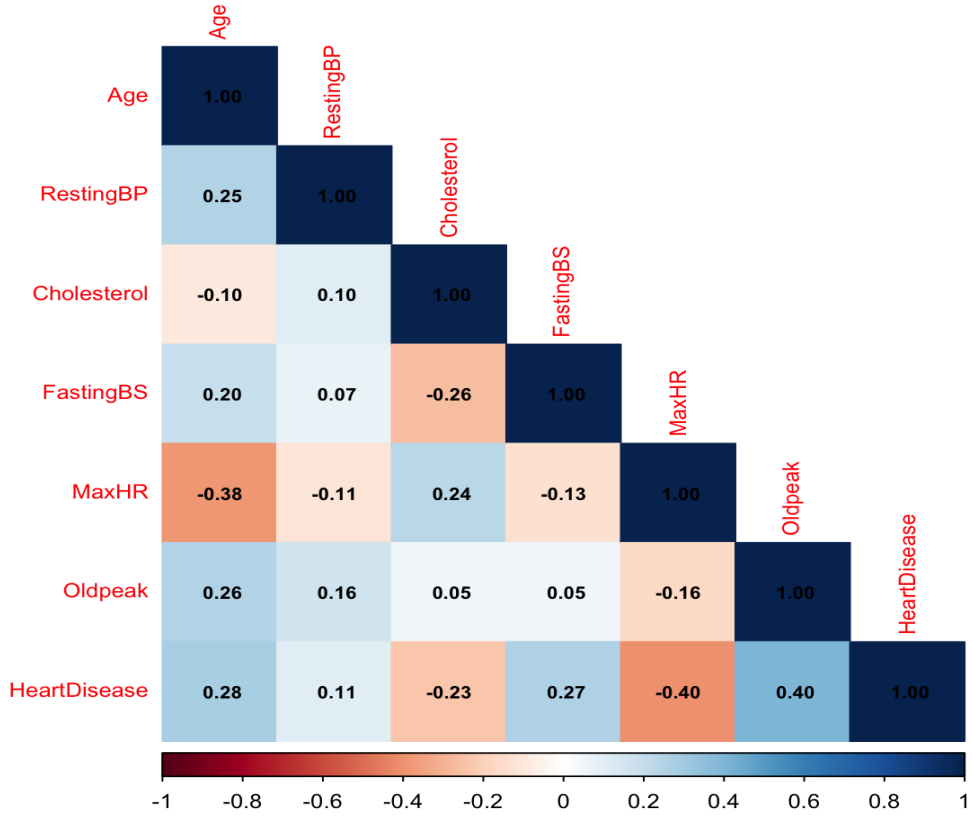



A tendência (linha azul) é **levemente negativa**, sugerindo que **quanto maior a idade, menor tende a ser o colesterol**.

# GRÁFICOS

# MATRIZ DE CORRELAÇÃO

02




**MaxHR vs Age**: **-0.38** → Correlação negativa moderada (esperado fisiologicamente)

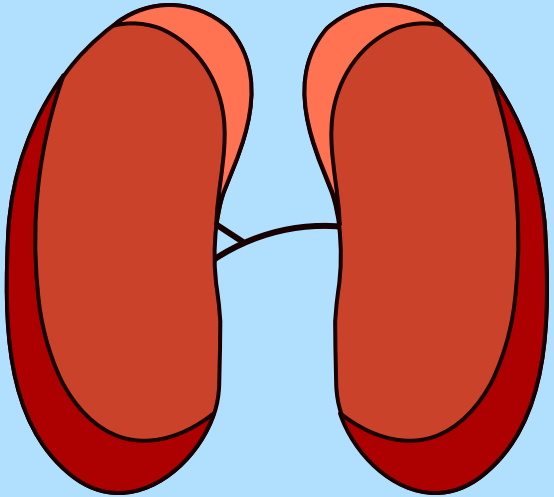
**MaxHR vs HeartDisease**: `→ Leve relação positiva (pode indicar esforço cardíaco)

**FastingBS vs HeartDisease**: **+0.27** → Glicemia em jejum elevada aparece como um fator de risco leve.

**RestingBP** → Correlações fracas com todas as variáveis, inclusive com HeartDisease.

03

# Testes de Normalidade



03

# Shapiro-Wilk

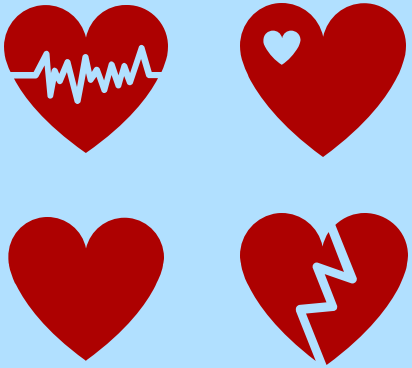
```
> # Teste de normalidade para Age
> shapiro.test(heart$Age)

        Shapiro-Wilk normality test

data:  heart$Age
W = 0.99101, p-value = 2.165e-05
```

Escolhemos esse em vez do Kolmogorov pelo tamanho do banco.

O teste de normalidade escolhido foi o Shapiro-Wilk que indicou que a variável Age não segue distribuição normal ($p < 0.05$). No entanto, como a regressão linear não exige normalidade da variável independente, apenas dos resíduos, o modelo pode ser aplicado normalmente.

04

# Coeficiente de Correlação



04

# Pearson

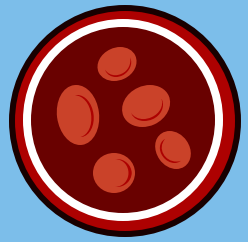
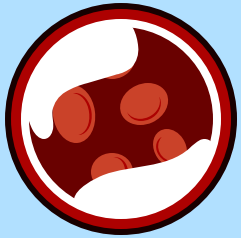
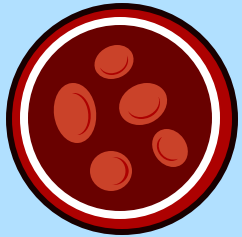
```
>
> # Correlação entre variáveis numéricas
> cor.test(heart$Age, heart$Cholesterol, method = "pearson")

        Pearson's product-moment correlation

data:  heart$Age and heart$Cholesterol
t = -2.8969, df = 916, p-value = 0.003858
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.15900537 -0.03076757
sample estimates:
        cor
-0.09528177
```

## Direção da correlação

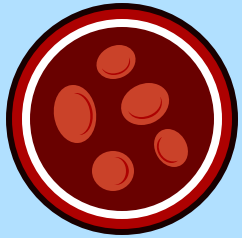
- A correlação é **negativa**: -0.095
- Isso indica que, **à medida que a idade aumenta, os níveis de colesterol tendem a diminuir levemente**.

## Força da correlação

- O valor de -0.095 indica uma **correlação muito fraca**.
- Correlações abaixo de |0.1| geralmente são consideradas **desprezíveis**.
- Ou seja, **idade praticamente não influencia o colesterol** neste conjunto de dados de forma relevante.

# Escolhendo variáveis para os modelos

# Regressão Linear – Previsão de Cholesterol

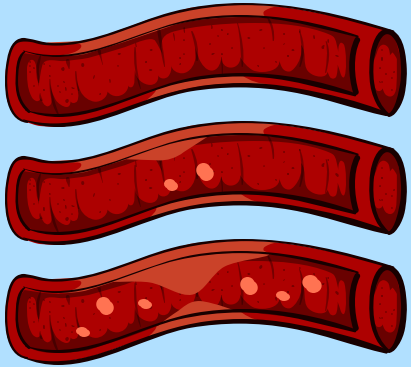
| Variável | Correlação com Cholesterol | Comentário |
|---|---|---|
| **Age** | **-0.10** | Correlação muito fraca, mas mantemos. |
| **MaxHR** | **+0.24** | Correlação fraca positiva — mantemos. |
| **ChestPainType** | Categórica | Pode capturar padrões clínicos importantes. |
| **RestingBP** | **+0.10** | Correlação fraca, muito dispersa — **não usar.** |
| **FastingBS** | **-0.26** | Leve correlação negativa, mas variável binária e com possível viés — **não usar.** |
| **Oldpeak** | **+0.05** | Correlação quase nula — **não usar.** |
| **HeartDisease** | **-0.23** | Pode enviesar o modelo — **não usar como preditora.** |

# Regressão Logística – Previsão de HeartDisease

| Variável | Correlação com HeartDisease | Comentário |
|---|---|---|
| **Age** | **+0.28** | Correlação fraca positiva — mantemos. |
| **MaxHR** | **-0.40** | Correlação negativa moderada — ótima variável. |
| **Cholesterol** | **-0.23** | Fraca, mas útil como variável complementar. |
| **ChestPainType** | Categórica | Fortemente associada clinicamente — mantemos. |
| **FastingBS** | **+0.27** | Binária, possível viés por medição — **não usar.** |
| **RestingBP** | **+0.11** | Muito fraca — **não usar.** |
| **Oldpeak** | **+0.40** | Boa correlação — mas foi deixada de fora por simplificação do modelo. |

05

# Regressão Linear



05

## Objetivo:

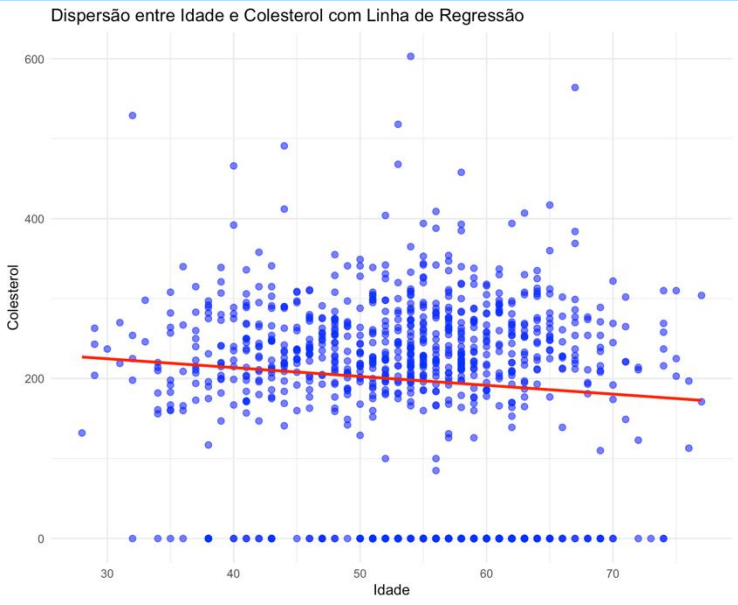
Prever o nível de colesterol com base em:

- Idade (Age)
- Frequência cardíaca máxima (MaxHR)
- Tipo de dor no peito (ChestPainType)

## Avaliação do Modelo:

```
> # Avaliação
> pred_lm <- predict(modelo_lm)
> residuos <- heart$Cholesterol - pred_lm
> mae <- mean(abs(residuos))
> rmse <- sqrt(mean(residuos^2))
> r2 <- summary(modelo_lm)$r.squared
> cat("MAE:", mae, "\nRMSE:", rmse, "\nR²:", r2)
MAE: 81.30562
RMSE: 105.7309
R²: 0.06466154>
```

## Gráfico de dispersão com linha de regressão:




## Retorno:

O valor de colesterol previsto e classificação do valor retornado

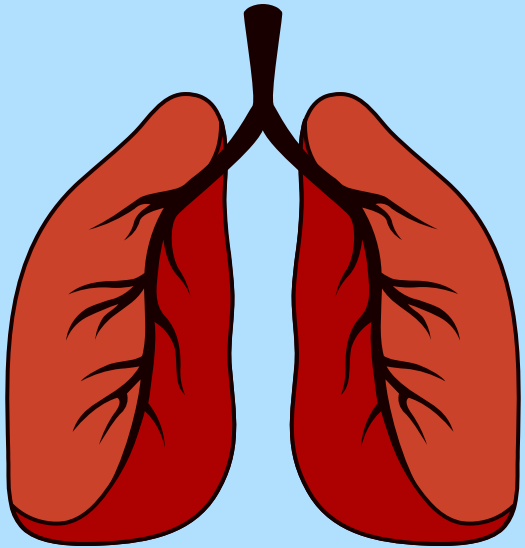
Abaixo de 200 = Desejável

200 a 239 = Limítrofe

240 ou mais = Alto

06

# Regressão Logística



06

## Objetivo:

Classificar se o paciente possui doença cardíaca com base em:

- Idade (Age)
- Frequência cardíaca máxima (MaxHR)
- Tipo de dor no peito (ChestPainType)
- Colesterol (Cholesterol)

## Retorno:

O valor de risco e probabilidade estimada de risco em % pelo modelo

Risco = 1 (Modelo prevê que o paciente **tem risco** de doença cardíaca)

Risco = 0 (Modelo prevê **sem risco** de doença cardíaca)

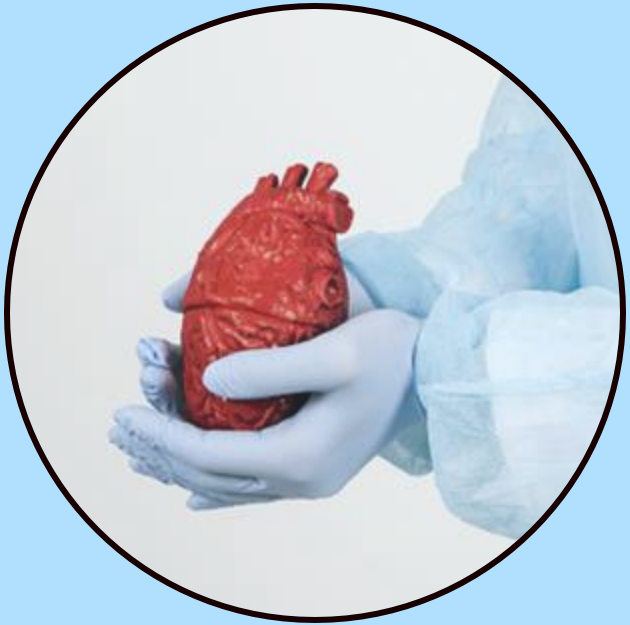
Probabilidade = Probabilidade estimada de risco (em %) pelo modelo

## Avaliaçäo do modelo:

```
> # Exibir matriz e métricas
> print(matriz)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 294  82
         1 116 426

               Accuracy : 0.7843
                 95% CI : (0.7563, 0.8105)
    No Information Rate : 0.5534
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5601

 Mcnemar's Test P-Value : 0.01902

            Sensitivity : 0.8386
            Specificity : 0.7171
         Pos Pred Value : 0.7860
         Neg Pred Value : 0.7819
             Prevalence : 0.5534
         Detection Rate : 0.4641
   Detection Prevalence : 0.5904
      Balanced Accuracy : 0.7778

       'Positive' Class : 1
```

07

# Conclusão



07

# Conclusão

Foram deixadas de fora variáveis com **baixa correlação**, **risco de redundância** ou que **não acrescentam poder preditivo significativo**.

Com isso o  projeto aplicou com sucesso técnicas de regressão linear e logística para prever o colesterol e o risco de doença cardíaca com base em dados clínicos. Os modelos apresentaram bom desempenho e foram integrados a uma API REST, permitindo previsões em tempo real. A abordagem demonstrou a aplicabilidade do machine learning em um tema de grande relevância social, unindo teoria estatística à prática em um cenário real.