Fatores de Risco para Doenças Cardíacas

Uma Análise Preditiva Utilizando Regressão Logística

Por: Caio Ribeiro de Carvalho e Thiago Henrique Moço Fonseca

Introdução e Objetivo

As doenças cardiovasculares são uma das principais causas de mortalidade global.

Este estudo aplica ciência de dados para identificar fatores de risco associados a doenças cardíacas.

Utilizamos regressão logística para construir um modelo preditivo eficiente.

Conjunto de Dados

Fonte

Dataset heart.csv da plataforma Kaggle com 918 pacientes (746 após pré-processamento).

Variáveis Clínicas

Idade, gênero, tipo de dor no peito, pressão sanguínea, colesterol, ECG, frequência cardíaca máxima.

Variável Alvo

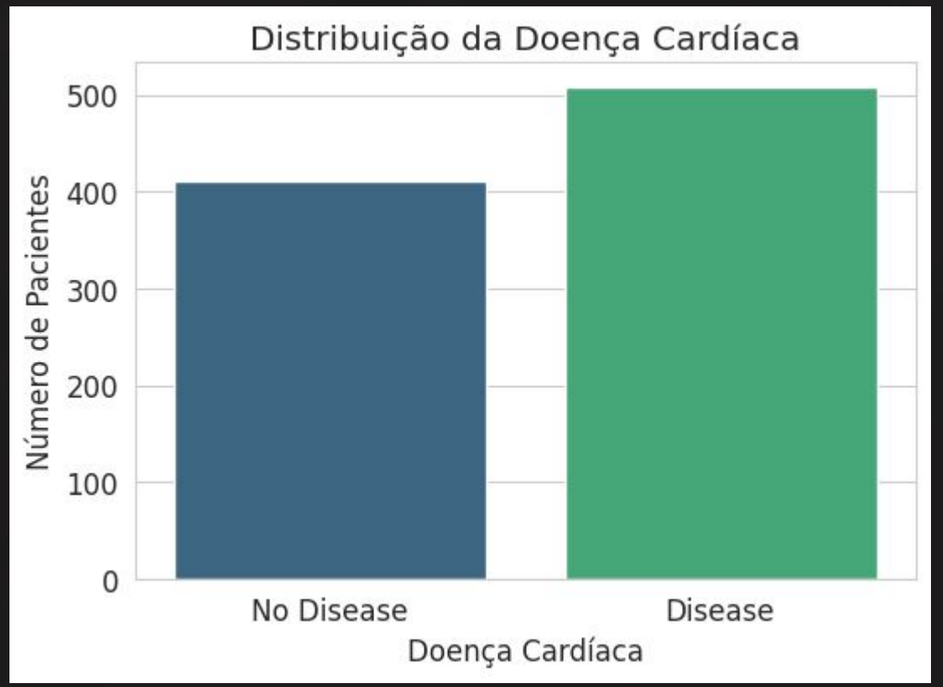
HeartDisease: presença (1) ou ausência (0) de doença cardíaca.

Distribuição da Variável Alvo antes do Processamento

O dataset apresenta uma distribuição equilibrada entre pacientes com e sem doença cardíaca:

- Sem doença (0): 410 pacientes (44,66%)
- Com doença (1): 508 pacientes (55,33%)

Esta distribuição balanceada favorece o treinamento do modelo preditivo.



Pré-processamento dos Dados

Verificação Inicial

Inspeção dos tipos de dados e estatísticas básicas.

Tratamento de Valores

Remoção de 172 linhas com valores zero em RestingBP e Cholesterol.

Codificação de Variáveis

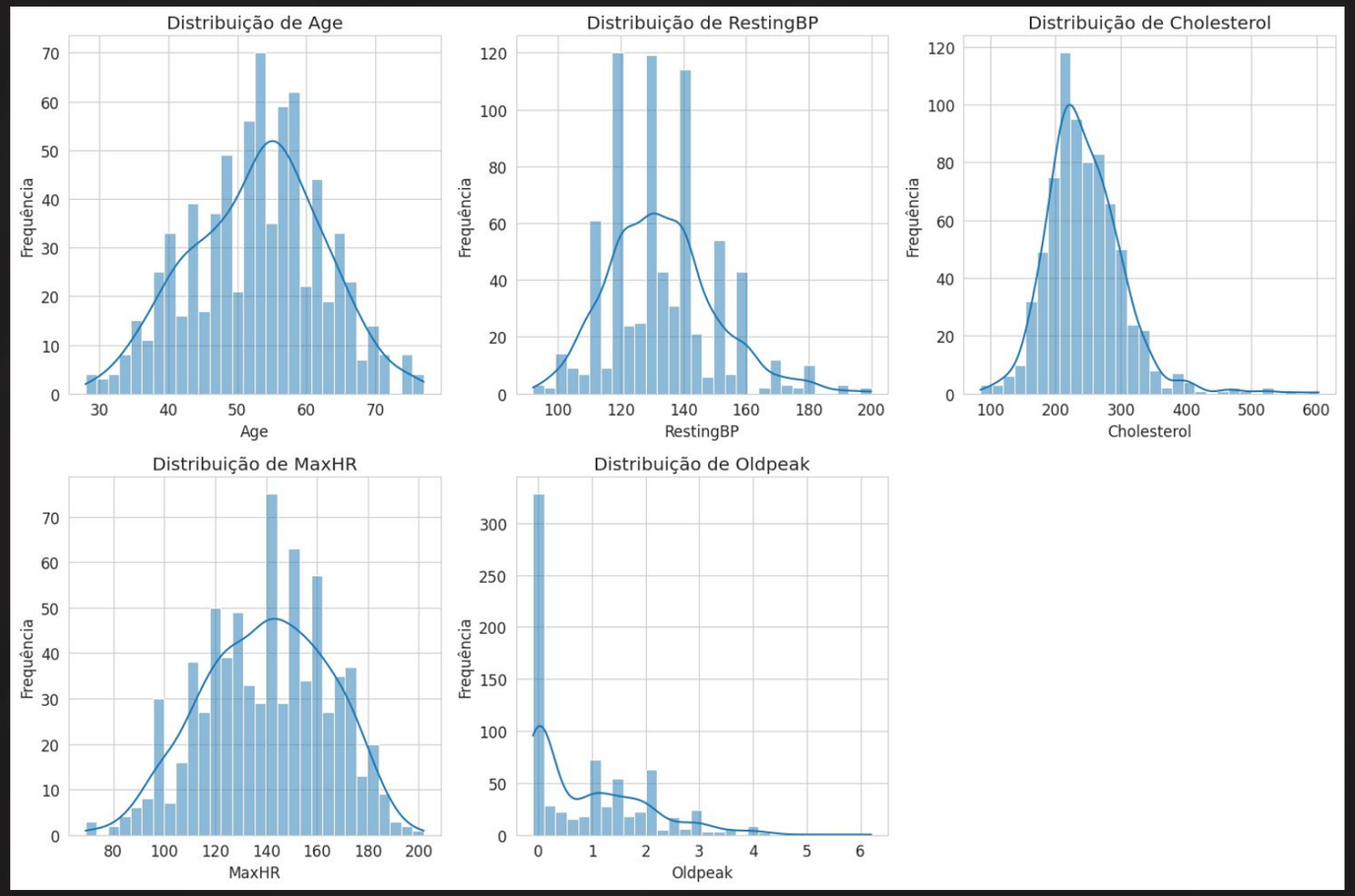
Conversão de variáveis categóricas usando One-Hot Encoding.

Análise Exploratória

Distribuição das Variáveis

A maioria das variáveis numéricas apresenta distribuição próxima à normal.

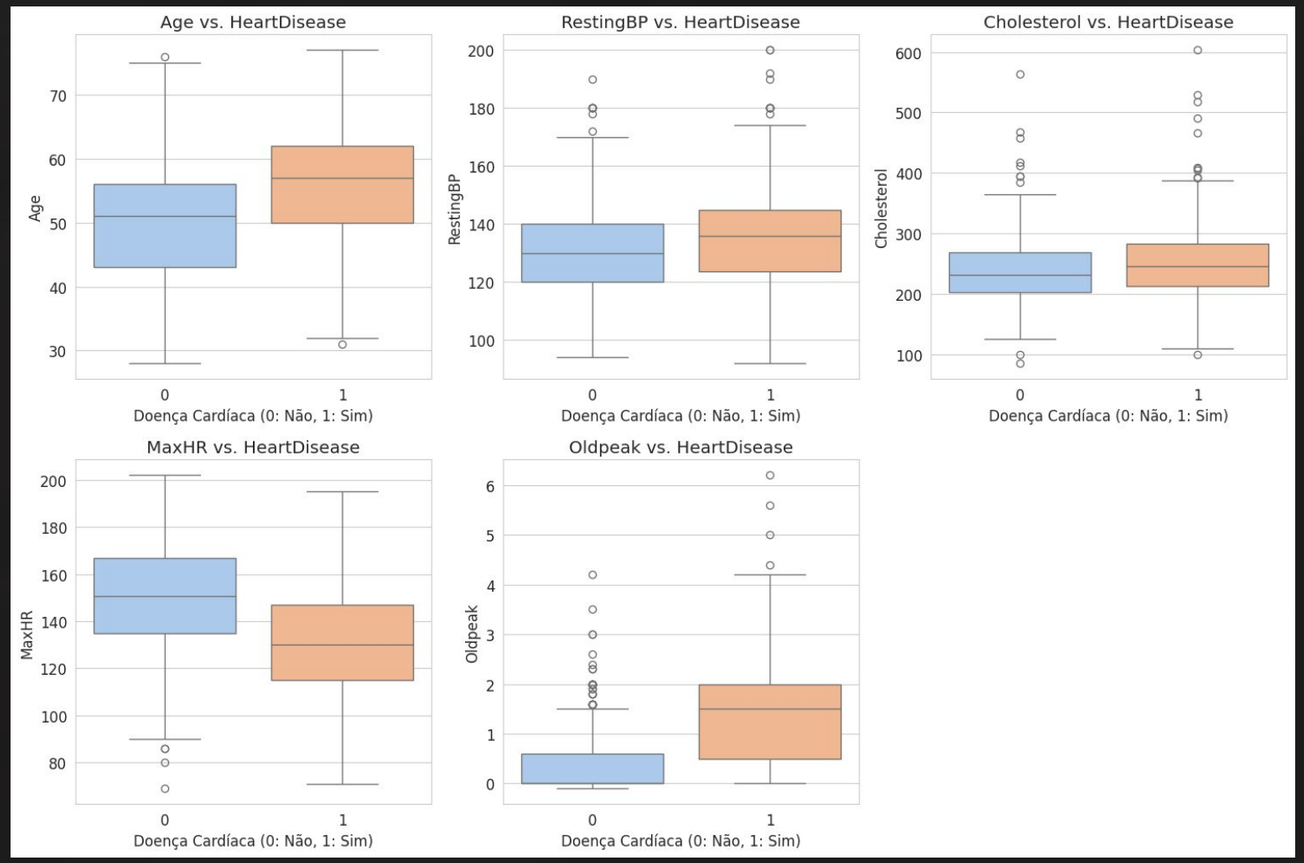
Dataset balanceado: 47,32% sem doença e 52,68% com doença cardíaca após o processamento de dados.



Relação entre Variáveis e Doença Cardíaca

Observações Importantes:

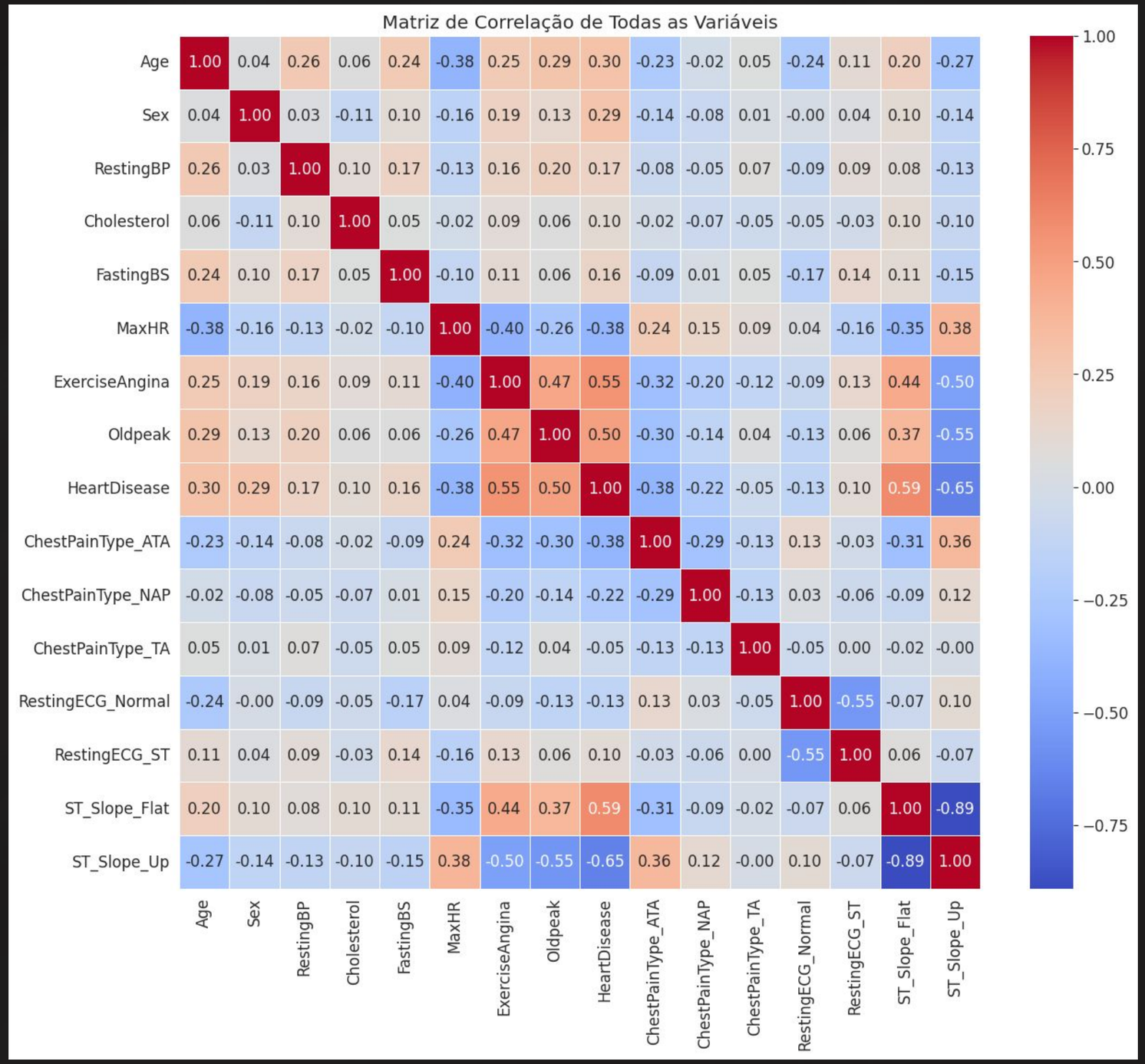
- Pacientes com doença tendem a ser mais velhos
- Frequência cardíaca máxima menor em pacientes com doença
- Valores de Oldpeak mais altos em pacientes com doença
- Homens apresentam maior prevalência de doença cardíaca



Correlação entre Variáveis

Principais correlações com doença cardíaca:

- ST Slope Flat: +0,59
- ExerciseAngina: +0,55
- Oldpeak: +0,50
- MaxHR: -0,38
- ChestPainType ATA: -0,38
- Age: +0,30



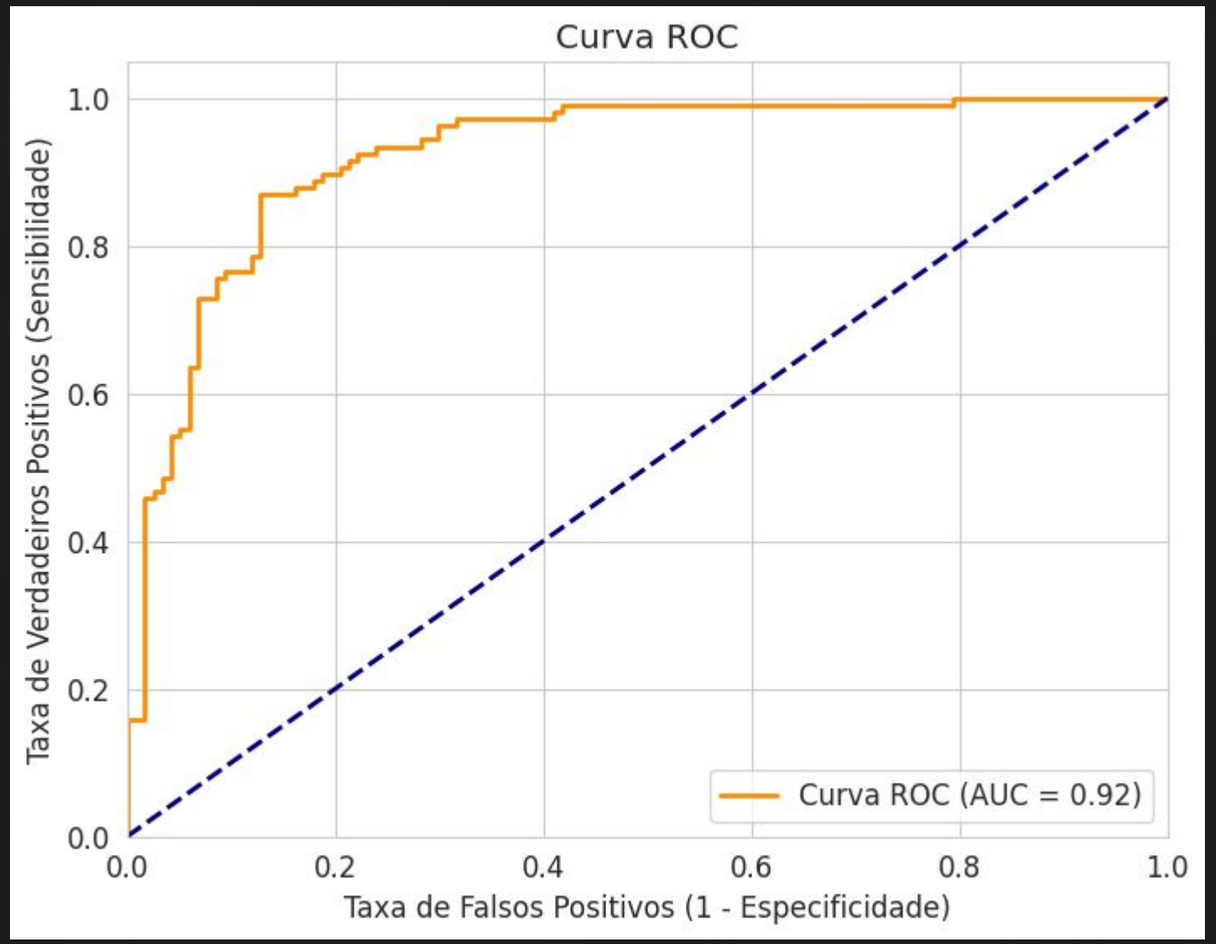
Modelo de Regressão Logística

Implementação do Modelo:

- Divisão: 70% treino, 30% teste
- Escalonamento das variáveis numéricas
- Pipeline para pré-processamento e modelagem

Métricas de Desempenho:

- Acurácia: 86,61%
- Precisão: 86,54%
- Recall: 89,36%
- F1-Score: 87,93%
- AUC ROC: 93,29%



Principais Fatores de Risco Identificados

100%

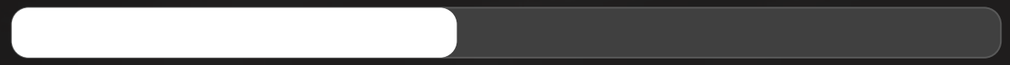
ST Slope Flat

Coeficiente: 2,45 - Principal preditor de doença cardíaca

63%

Dor no Peito Assintomática

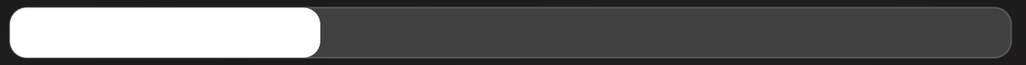
Coeficiente: 1,54 - Segundo preditor mais forte



45%

Angina Induzida por Exercício

Coeficiente: 1,10 - Aumenta substancialmente o risco



31%

Oldpeak (Depressão ST)

Coeficiente: 0,77 - Forte associação com doença cardíaca