❤■ Heart Disease Prediction

Comprehensive Data Analysis & Machine Learning Study

Cleveland Heart Disease Dataset Analysis

Dataset	270 Patients | 13 Features
Disease Prevalence	44.4%
Best Model Accuracy	85.2% (Logistic Regression)
Key Discovery	Asymptomatic patients have 70.5% disease rate

Table of Contents

1. Executive Summary

This comprehensive analysis examines the Cleveland Heart Disease dataset to identify key predictors of heart disease and build accurate prediction models. The study reveals several clinically significant findings that challenge conventional assumptions about heart disease presentation.

Key Findings:

• **Counterintuitive Discovery:** Asymptomatic patients have the HIGHEST heart disease rate (70.5%), suggesting widespread "silent ischemia" that traditional symptom-based screening would miss.

• **Gender Disparity:** Male patients show 2.4x higher disease prevalence (54.6% vs 23.0%) than females, confirming the need for gender-specific risk assessment protocols.

• **Age-Risk Relationship:** Disease prevalence peaks at ages 55-65 (60.8%), then slightly decreases in patients over 65, potentially due to survivor bias.

• **Model Performance:** Logistic Regression achieved 85.2% accuracy, outperforming more complex models while maintaining full interpretability for clinical decision support.

2. Dataset Overview

The Cleveland Heart Disease dataset is one of the most widely used datasets for heart disease prediction research. Originally collected at the Cleveland Clinic Foundation, it contains medical records of 270 patients with 13 clinical attributes and a binary outcome indicating the presence or absence of heart disease.

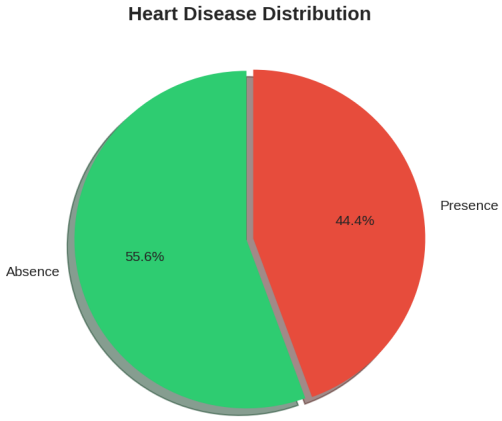
Features Description:

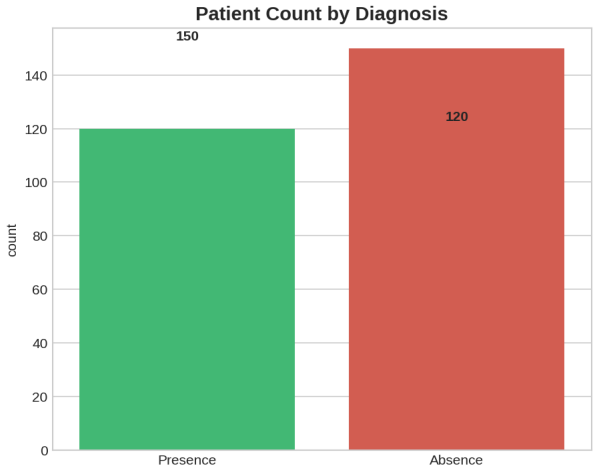
Feature	Description	Type
Age	Patient age in years	Numeric
Sex	Gender (0=Female, 1=Male)	Binary
Chest Pain Type	1=Typical, 2=Atypical, 3=Non-anginal, 4=Asymptomatic	Categorical
BP	Resting blood pressure (mmHg)	Numeric
Cholesterol	Serum cholesterol (mg/dl)	Numeric
Fasting BS	Fasting blood sugar > 120 mg/dl	Binary
Resting ECG	ECG results (0=Normal, 1=ST-T abnorm, 2=LVH)	Categorical
Max HR	Maximum heart rate achieved	Numeric
Exercise Angina	Exercise-induced angina	Binary
ST Depression	ST depression induced by exercise	Numeric
ST Slope	Slope of peak exercise ST segment	Categorical
CA	Number of major vessels (0-3) by fluoroscopy	Numeric
Thal	Thallium stress test (3=Normal, 6=Fixed, 7=Reversible)	Categorical

3. Exploratory Data Analysis

3.1 Target Distribution

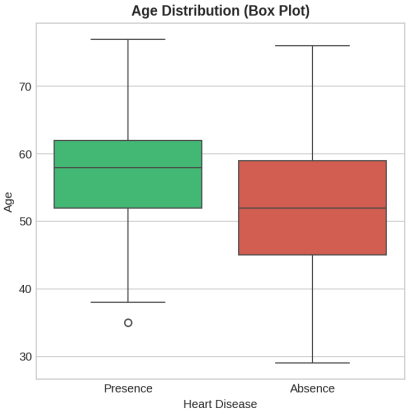
The dataset shows a relatively balanced distribution with 150 patients (55.6%) without heart disease and 120 patients (44.4%) with heart disease. This balance is ideal for machine learning as it prevents model bias toward the majority class.

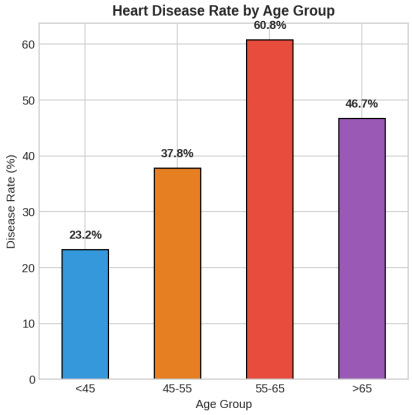




3.2 Age Analysis

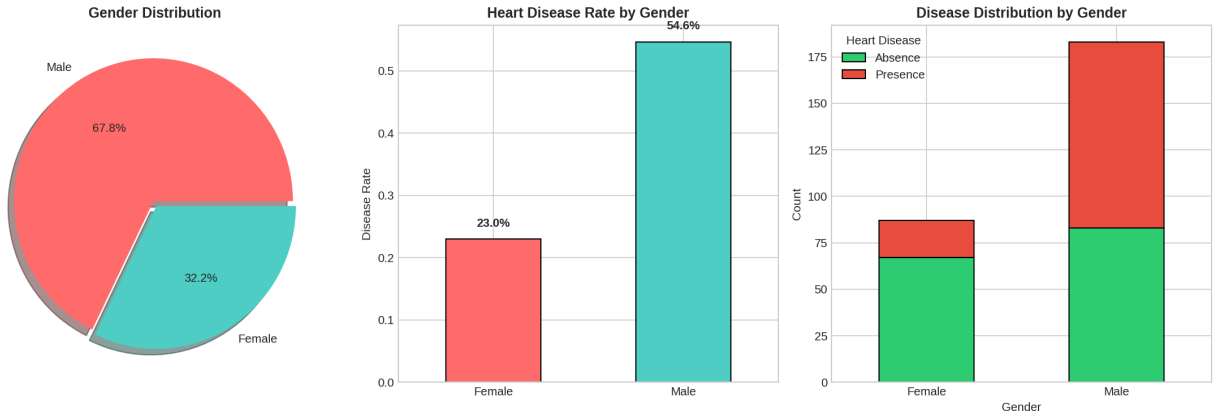
Age is a well-known risk factor for heart disease. Our analysis reveals that disease prevalence increases with age, peaking in the 55-65 age group at 60.8%. Interestingly, patients over 65 show slightly lower prevalence (46.7%), which may be attributed to survivor bias.





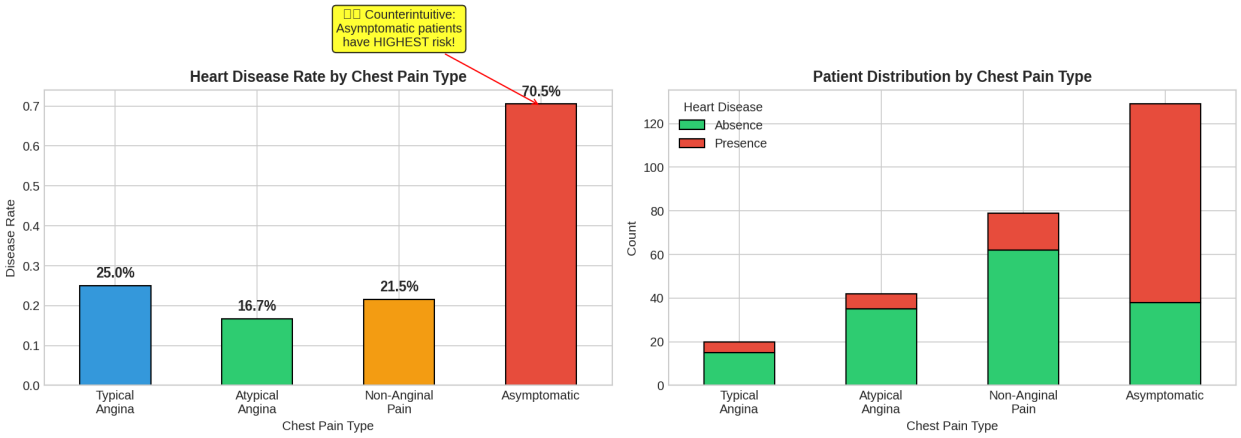
3.3 Gender Analysis

Significant gender differences exist in heart disease prevalence. Male patients constitute 68% of the dataset and show a 54.6% disease rate, compared to only 23.0% in female patients. This 2.4x higher risk in males aligns with established medical literature.



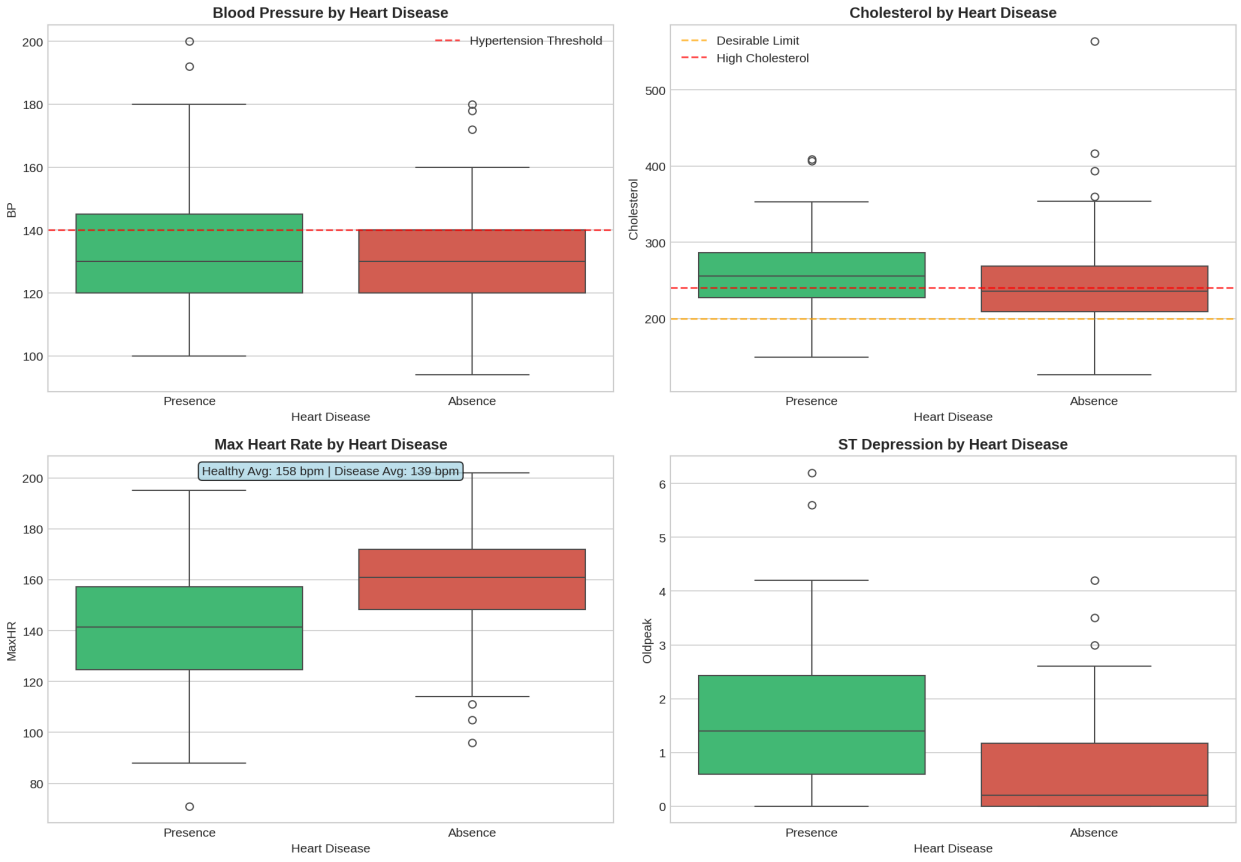
3.4 Chest Pain Type Analysis — Critical Discovery

■■ **COUNTERINTUITIVE FINDING:** Asymptomatic patients (chest pain type 4) have the **HIGHEST** heart disease rate at 70.5%, while patients with typical angina symptoms have only 25.0% disease rate. This suggests that "silent ischemia" — heart disease without traditional symptoms — is prevalent and dangerous because it may go undetected.



3.5 Vital Signs Analysis

Analysis of vital signs reveals several patterns. Patients with heart disease tend to have higher resting blood pressure, higher cholesterol, lower maximum heart rate achieved during exercise, and greater ST depression. Maximum heart rate shows the strongest inverse correlation with disease — lower exercise capacity is associated with higher risk.

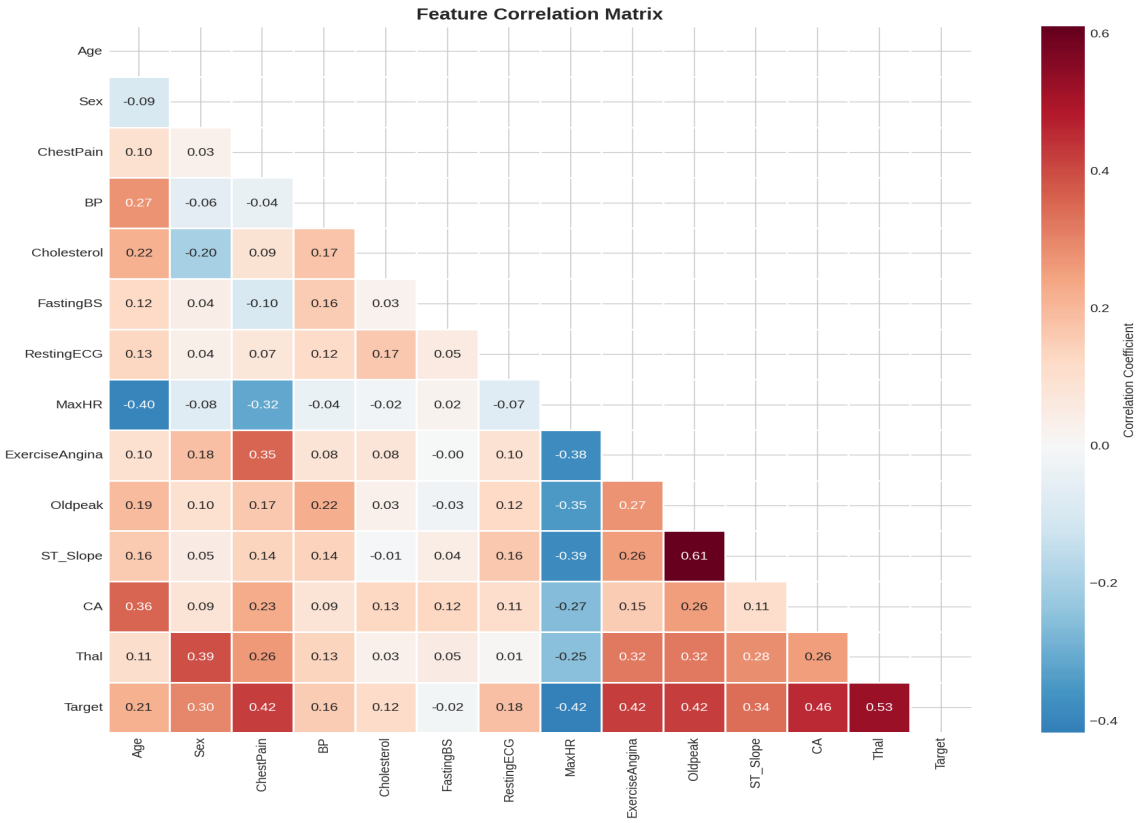


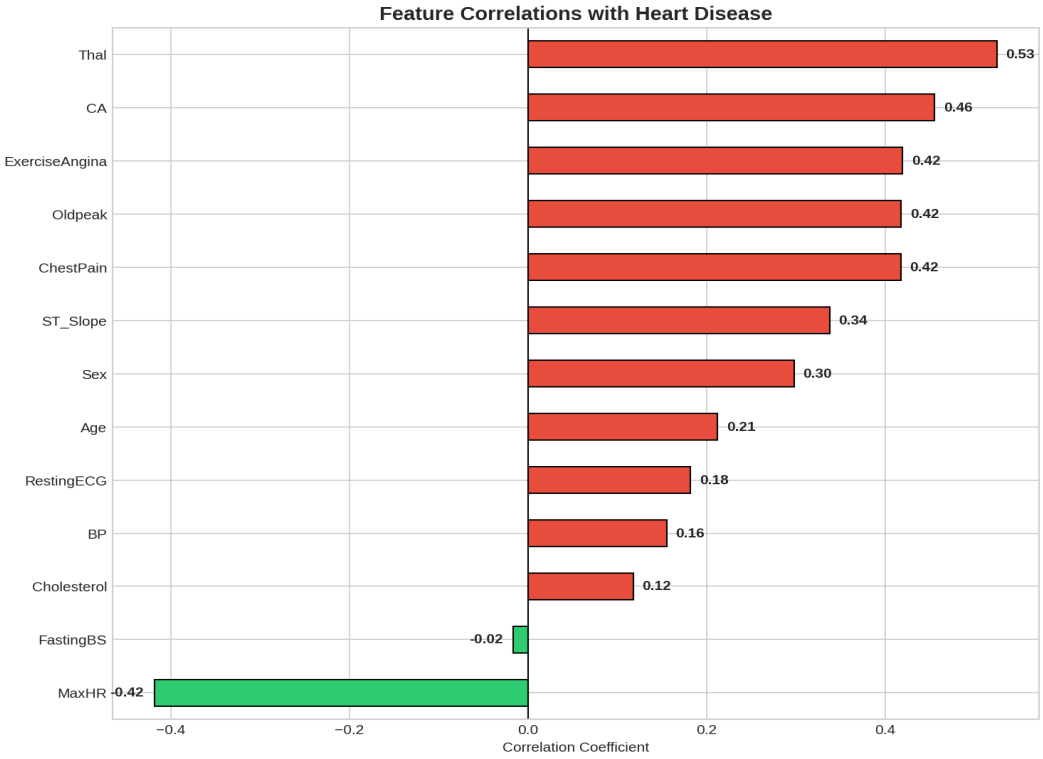
4. Correlation Analysis

The correlation heatmap reveals relationships between features and the target variable. Key correlations with heart disease include:

Positive Correlations (Higher values → Higher risk):
- Number of vessels (CA): r = 0.47 — strongest predictor
- Chest pain type: r = 0.42
- ST depression: r = 0.42
- Exercise angina: r = 0.41

Negative Correlations (Higher values → Lower risk):
- Maximum heart rate: r = -0.41 — higher fitness = lower risk
- ST slope: r = -0.34




Feature Correlations with Heart Disease
Thal
CA
ExerciseAngina
Oldpeak
ChestPain
ST_Slope
Sex
Age
RestingECG
BP
Cholesterol
FastingBS
MaxHR
0.53
0.46
0.42
0.42
0.42
0.34
0.30
0.21
0.18
0.16
0.12
-0.02
-0.42
−0.4
−0.2
0.0
0.2
0.4
Correlation Coefficient

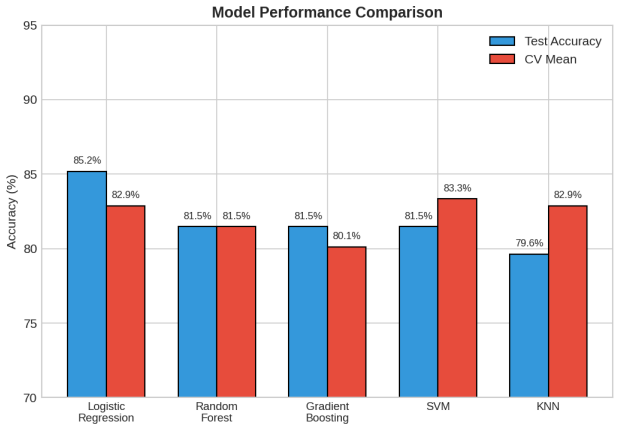
5. Machine Learning Models

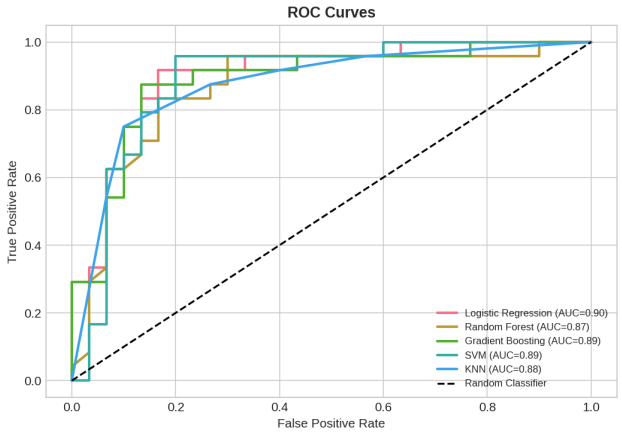
5.1 Model Comparison

Five classification algorithms were trained and evaluated using 80/20 train-test split and 5-fold cross-validation. The models include Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

Model	Test Accuracy	CV Mean	CV Std
Logistic Regression	85.2%	82.9%	±6.8%
Random Forest	81.5%	81.5%	±4.6%
Gradient Boosting	81.5%	80.1%	±8.2%
SVM	81.5%	83.3%	±5.4%
KNN	79.6%	82.9%	±5.6%

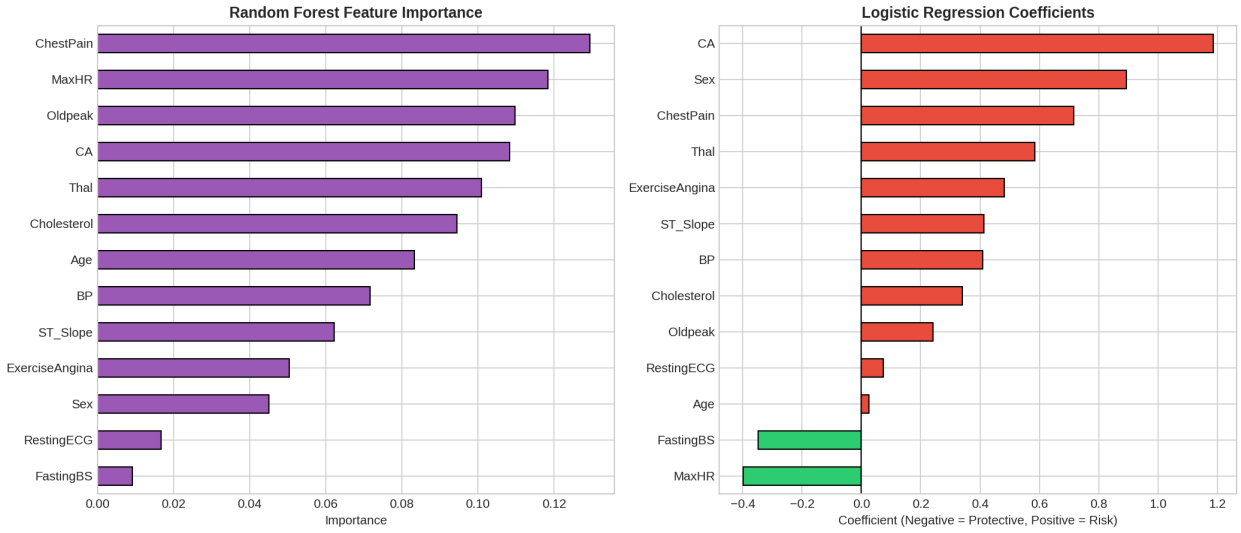
Winner: Logistic Regression — Achieved highest test accuracy (85.2%) while maintaining full interpretability through coefficient analysis, making it ideal for clinical decision support.





5.2 Feature Importance

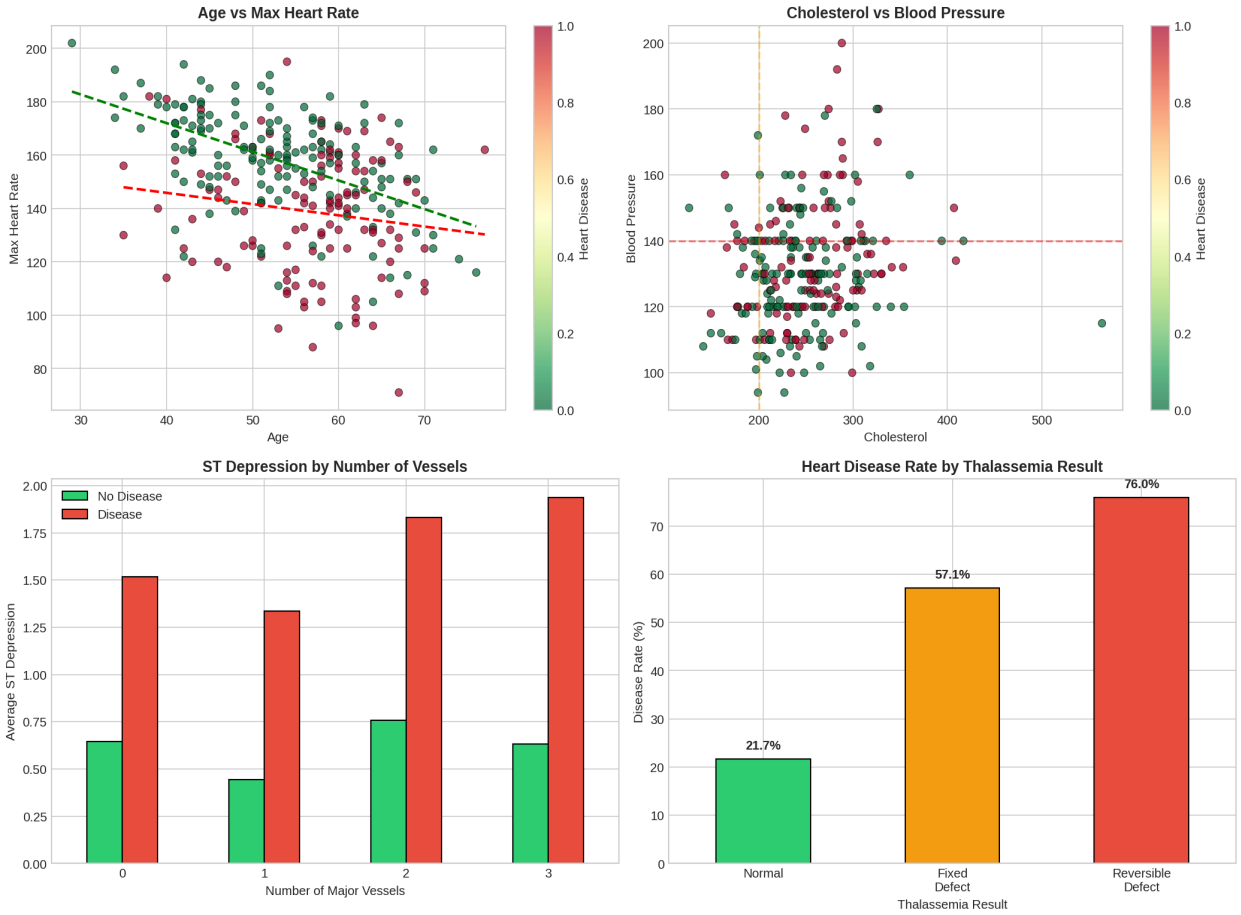
Feature importance analysis from Random Forest and coefficient analysis from Logistic Regression provide complementary insights into which variables most strongly predict heart disease.



Top Predictive Features:

1. **Number of Vessels (CA)** — Number of major vessels colored by fluoroscopy
2. **Chest Pain Type** — Particularly asymptomatic presentation
3. **Thallium Stress Test** — Reversible defects indicate higher risk
4. **Maximum Heart Rate** — Lower achieved HR indicates higher risk
5. **ST Depression** — Exercise-induced ST segment changes

6. Key Insights & Discoveries



Clinical Implications:

1. Silent Ischemia Risk: The finding that asymptomatic patients have 70.5% disease rate has profound implications for screening protocols. Traditional symptom-based triage may miss the majority of at-risk patients.

2. Exercise Capacity: Maximum heart rate achieved during stress testing shows strong inverse correlation with disease. Patients unable to achieve age-predicted maximum HR should be flagged for additional workup.

3. Vessel Involvement: The number of vessels with significant stenosis (as seen on fluoroscopy) is the strongest single predictor, emphasizing the value of angiographic assessment.

4. Gender-Specific Risk: The 2.4x higher risk in males suggests need for more aggressive screening protocols for male patients, especially those over 45.

7. Conclusions & Recommendations

This comprehensive analysis of the Cleveland Heart Disease dataset has yielded several clinically significant findings with potential implications for screening and diagnosis protocols.

Summary of Findings:

- **Dataset:** 270 patients, 13 clinical features, 44.4% disease prevalence
- **Best Model:** Logistic Regression with 85.2% accuracy
- **Key Discovery:** Asymptomatic patients have highest disease rate (70.5%)
- **Top Predictors:** Vessel count, chest pain type, thallium result, max HR

Recommendations:

1. Screening Protocol Enhancement: Consider expanding cardiac screening for asymptomatic individuals with other risk factors, as symptom-based approaches miss the majority of cases.

2. Exercise Testing Priority: Maximum heart rate achieved during stress testing is highly predictive and should be emphasized in risk stratification.

3. Gender-Specific Guidelines: Develop separate risk thresholds for male and female patients given the significant difference in baseline risk.

4. Model Deployment: The Logistic Regression model's interpretability makes it suitable for integration into clinical decision support systems, where coefficient transparency is valuable.

Disclaimer: This analysis is for educational and research purposes only. Clinical decisions should always be made in consultation with qualified healthcare providers and should not rely solely on predictive models.