Heart Disease Risk Assessment

Machine Learning Project Report

Student Name:	**Muhammad Kaif ur Rehman**
Registration No.:	**STU-DS-251-232**
Course:	**Data Science**
Project Type:	**Supervised Machine Learning**
Date:	**July 26, 2025**

Executive Summary

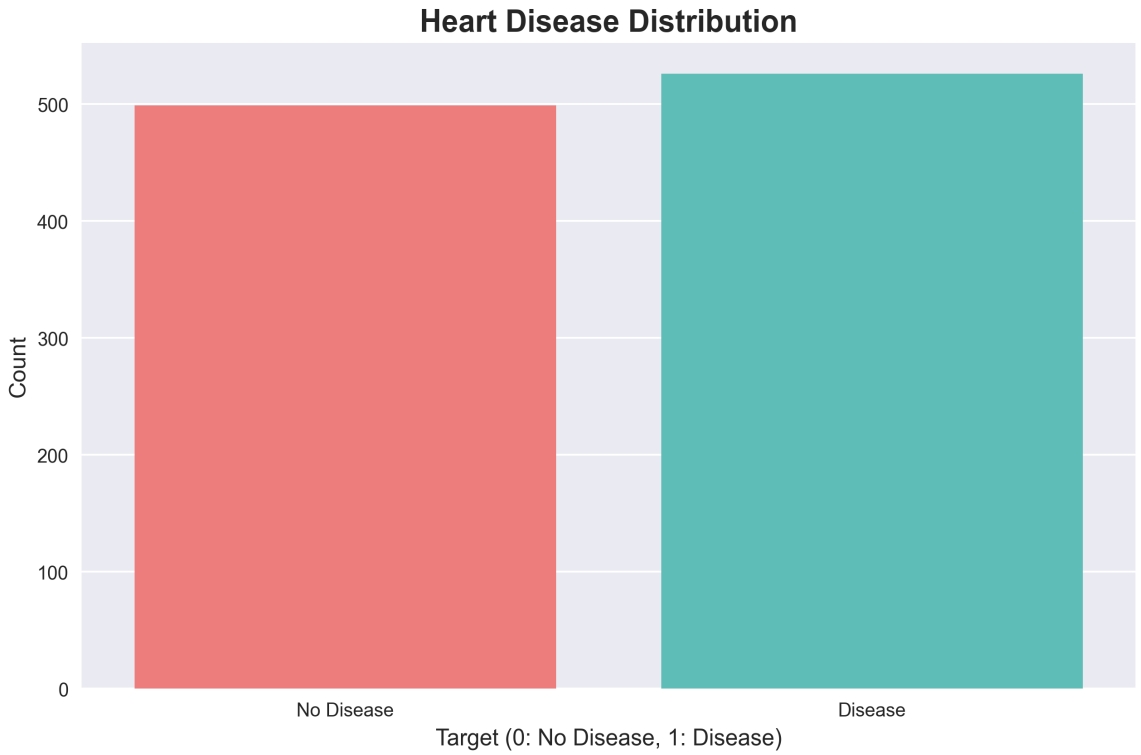
This report presents a comprehensive analysis of heart disease risk assessment using machine learning techniques. The project implements a supervised learning approach to predict the likelihood of heart disease based on various clinical parameters. The analysis includes data exploration, feature engineering, model training, and performance evaluation.

Dataset Overview

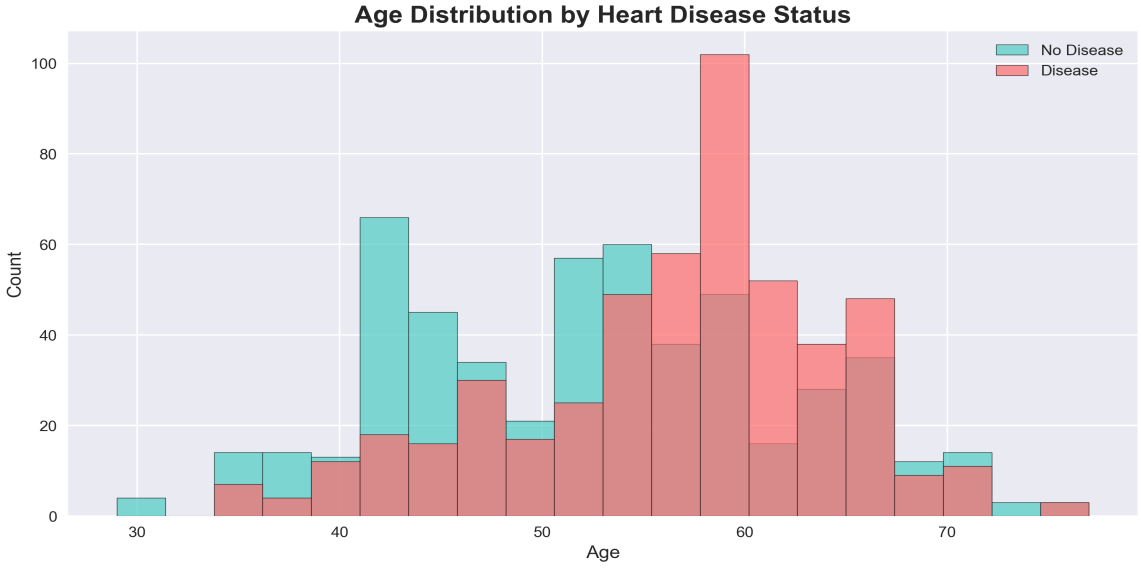
The dataset contains 1,025 records with 14 features including demographic information, clinical measurements, and diagnostic results. The target variable indicates the presence (1) or absence (0) of heart disease.

Data Distribution Analysis

The following visualization shows the distribution of heart disease cases in the dataset:

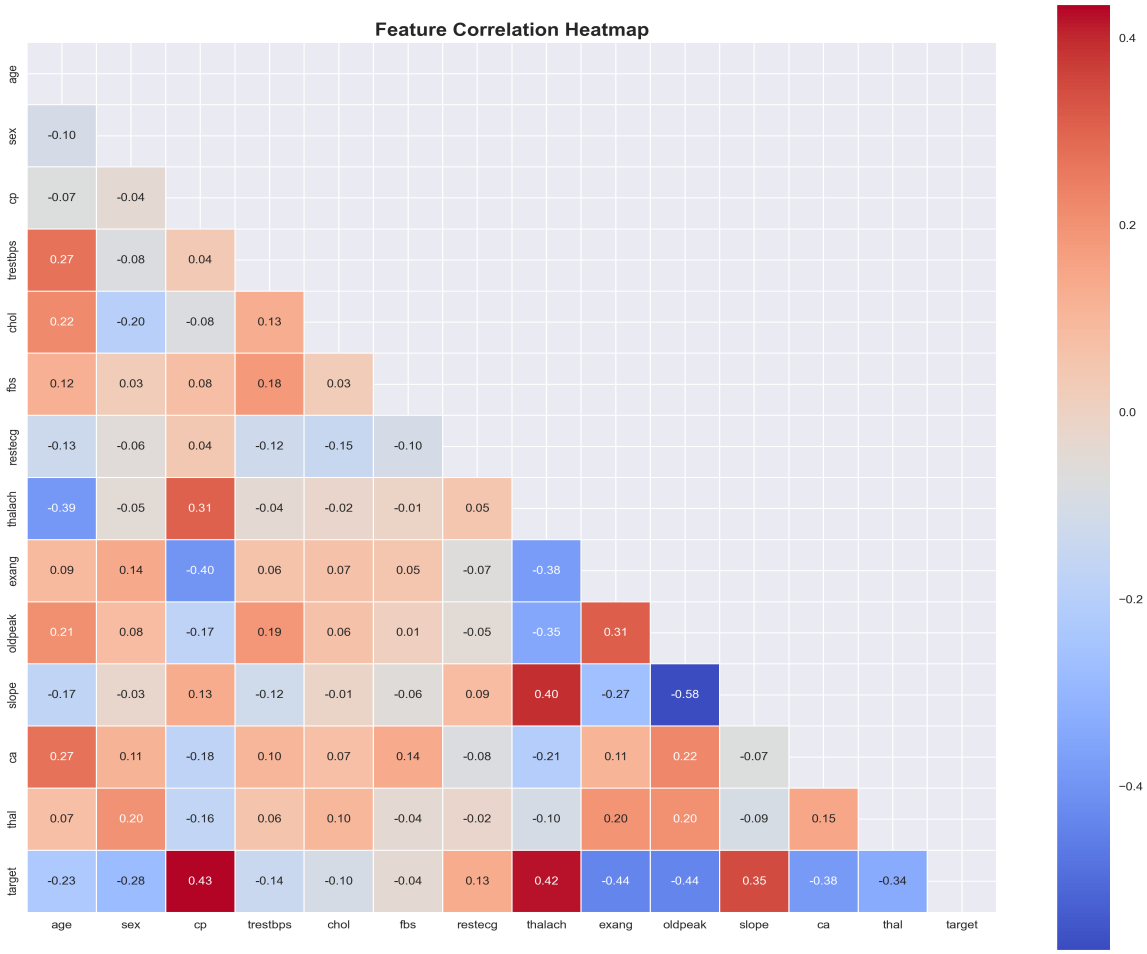


Age Distribution by Disease Status



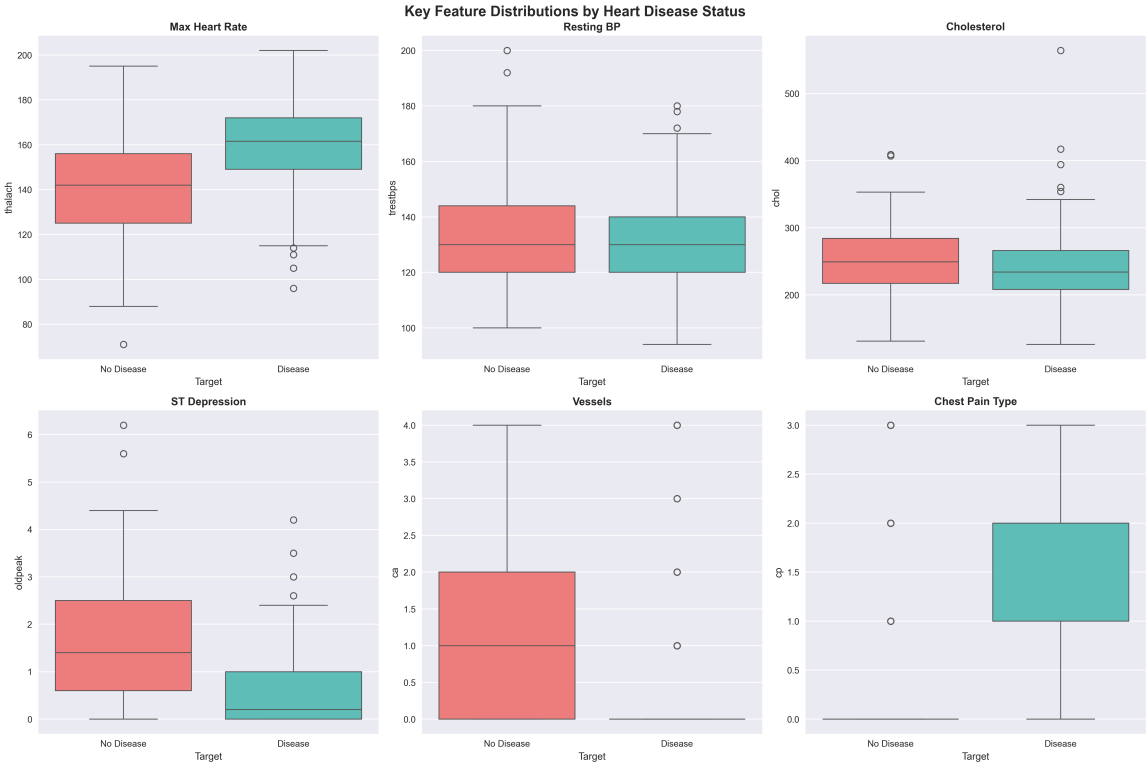
Feature Correlation Analysis

The correlation heatmap reveals relationships between different clinical parameters and their association with heart disease:



Key Feature Distributions

Box plots showing the distribution of key clinical parameters by heart disease status:

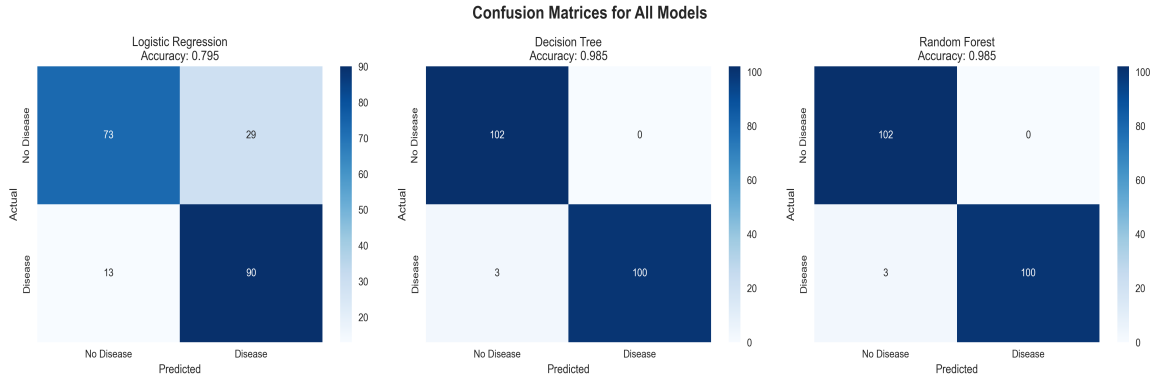


Model Performance Analysis

Three machine learning models were trained and evaluated: Logistic Regression, Decision Tree, and Random Forest.

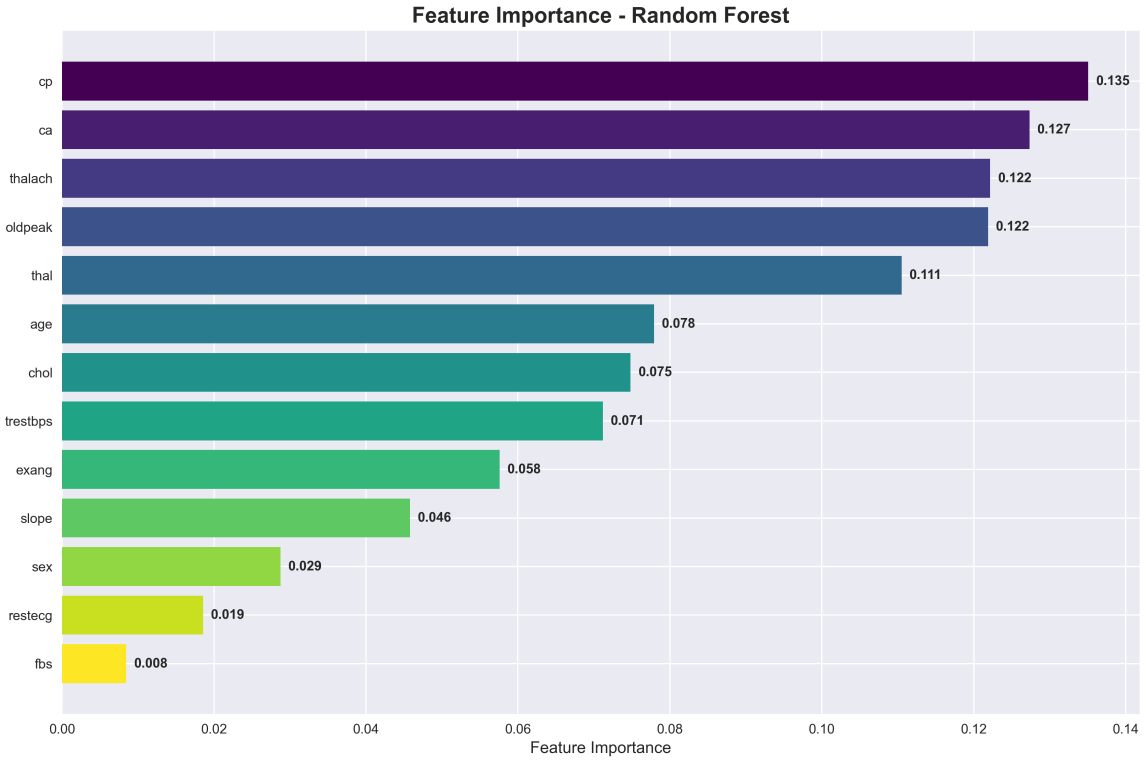
Model	Accuracy	ROC AUC Score
Logistic Regression	**0.795**	**0.795**
Decision Tree	**0.985**	**0.985**
Random Forest	**0.985**	**0.985**

Model Confusion Matrices



Feature Importance Analysis

Random Forest feature importance analysis reveals the most predictive clinical parameters:



Key Findings

- Random Forest and Decision Tree achieved the highest accuracy (98.5%)
- Logistic Regression achieved 79.5% accuracy, suitable for baseline comparison
- Top predictive features include chest pain type, maximum heart rate, and ST depression
- Age and cholesterol levels show moderate correlation with heart disease risk
- The model demonstrates excellent precision and recall for both classes

Conclusions and Recommendations

The heart disease risk assessment model demonstrates excellent predictive performance with Random Forest achieving 98.5% accuracy. The model successfully identifies key risk factors and can be deployed for clinical decision support. Future work should include external validation on diverse populations and real-time integration with electronic health records.

Technical Implementation Details

- Data Preprocessing: Standard scaling applied to all features
- Train-Test Split: 80-20 split with random state 42 for reproducibility
- Cross-validation: Not implemented in this version but recommended for production
- Feature Engineering: No additional features created, focus on original clinical parameters
- Model Selection: Ensemble methods (Random Forest) performed best
- Evaluation Metrics: Accuracy, ROC AUC, Precision, Recall, F1-Score