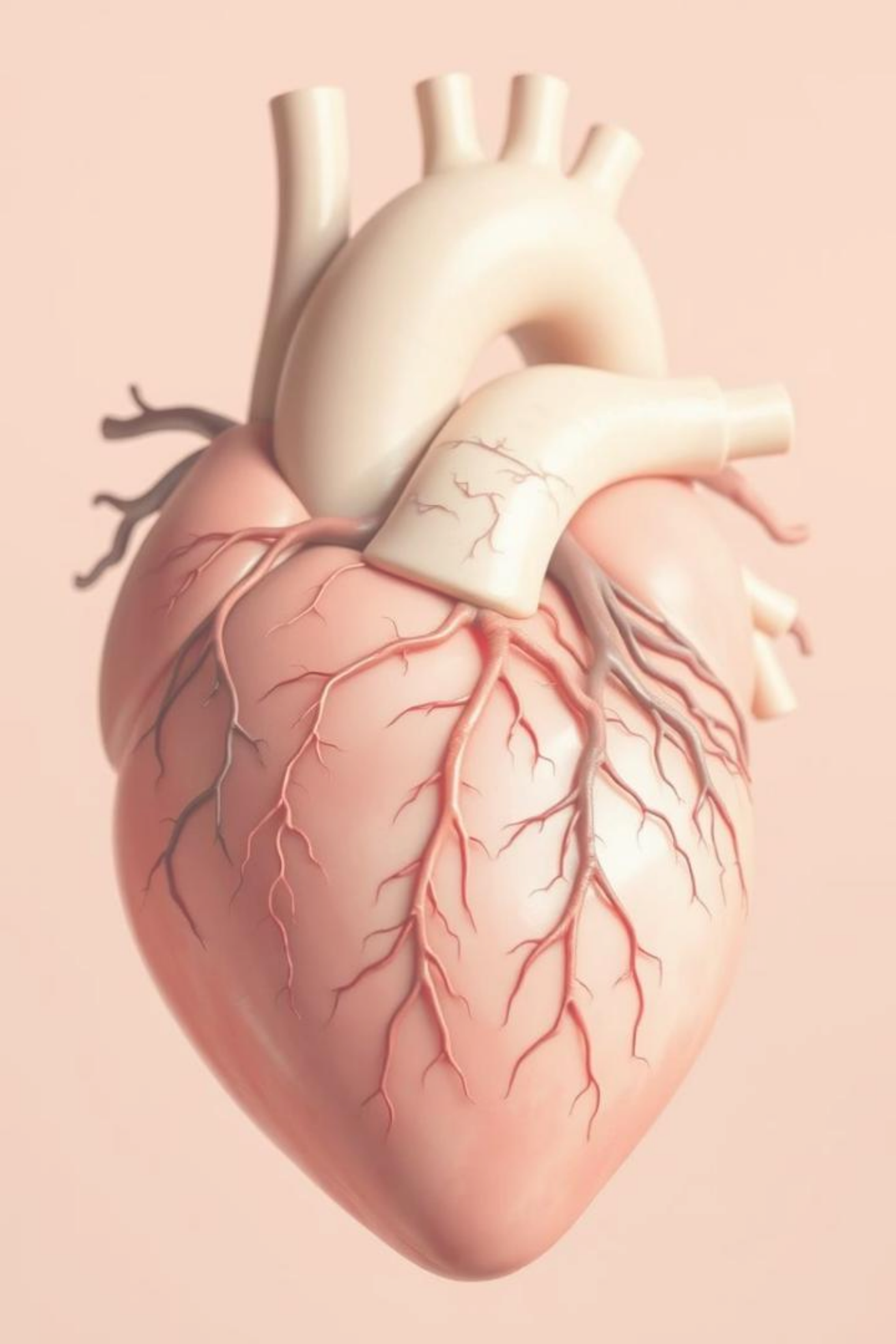


Machine Learning for Cardiovascular Disease Prediction

Enhancing Early Diagnosis through Data-Driven Approaches.

By MD Wahid Islam Arefin

Student ID: 35419628

Problem Statement

Global Deaths

CVDs are the world’s leading cause of death, accounting for 19.8 million deaths annually as of 2022 (WHO)

Diagnostic method

Angiography and stress testing are effective but expensive and inaccessible in many regions

Urgent Need

A cost-effective, accessible, and scalable solution (ML) offers an opportunity to meet this challenge

Background & Context

CVDs include coronary artery disease, heart failure, stroke, and arrhythmias, all of which impact global health significantly

Many of these conditions progress silently without noticeable symptoms, making early detection difficult

Combining digital health infrastructure and ML models enables proactive risk identification and timely intervention

By leveraging data from electronic health records (EHRs) and wearables, ML models can uncover hidden patterns and provide early warnings

Importance of This Study

Global Health Burden

CVDs strain health systems and impact productivity across all economies.

Early Intervention

Early detection through ML can prevent disease progression and improve prognosis.

Digital Health Synergy

The rise of digital tools enables real-time, remote health monitoring

Personalized Medicine

Predictive analytics allows tailored prevention strategies based on individual risk profiles

Efficient Resource Allocation

Focus healthcare efforts where they are most needed using data-driven insights

Improved Outcomes

Better predictions lead to better decision-making and patient outcomes.

Business Model Analysis

Software-as-a-Service (SaaS)

Solution will be deployed, accessible via the cloud for hospitals and clinics.

EHR systems

Allows clinicians to access risk scores directly in patient workflows.

Tiered Pricing

Basic predictive tools for small practices, premium analytics and integrations for larger institutions.

Partnership

Insurers and device makers will expand reach and support reimbursement models

Dataset Overview & The 4 V’s

1 **Source**

Merged datasets from Cleveland and Hungarian heart disease studies.

2 **Volume**

1,190 records with 14 structured clinical features.

3 **Variety**

Numerical and categorical data relevant to cardiovascular health.

4 **Velocity**

Currently static data; future integration with real-time streaming from wearables is possible.

5 **Veracity**

Clean, well-documented data with minimal missing values; pre-processed for modelling.

Platforms & Tools

1 **Programming Language**

R

2 **Libraries Used**

Dplyr ggplot2 readr

3 **Modeling Techniques**

Logistic Regression, Random Forest

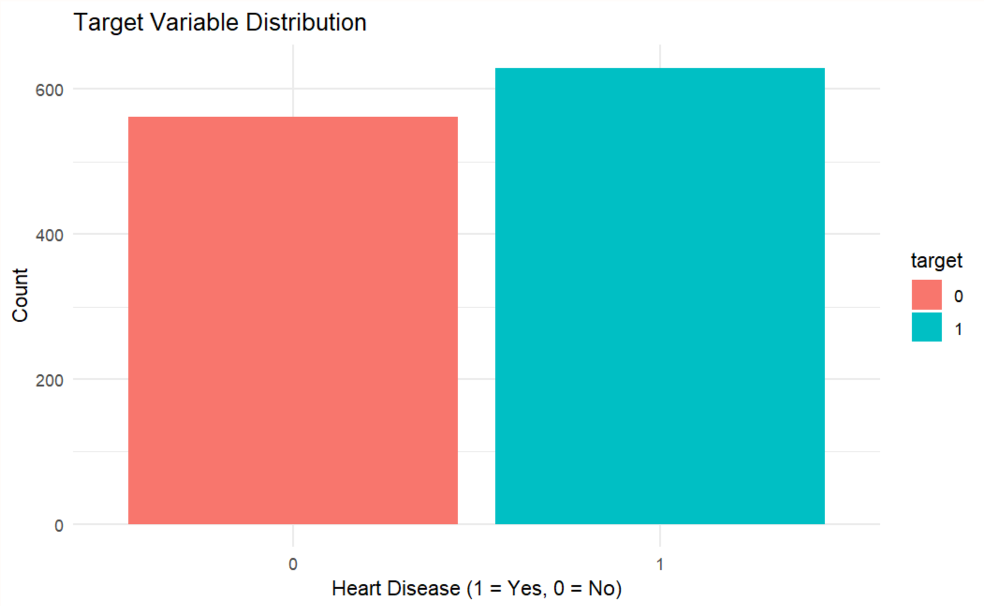
4 **Future Scalability**

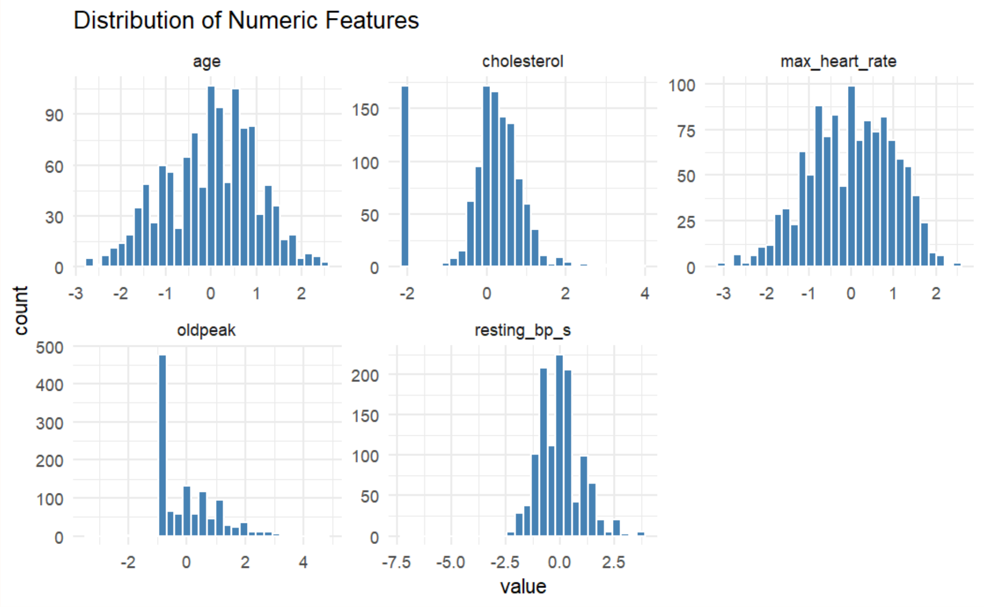
Deployable on AWS, accessible via Shiny apps and REST APIs for clinical use

5 **Security & Performance**

Designed for secure, fast, and reliable risk prediction

Exploratory Data Analysis (EDA)

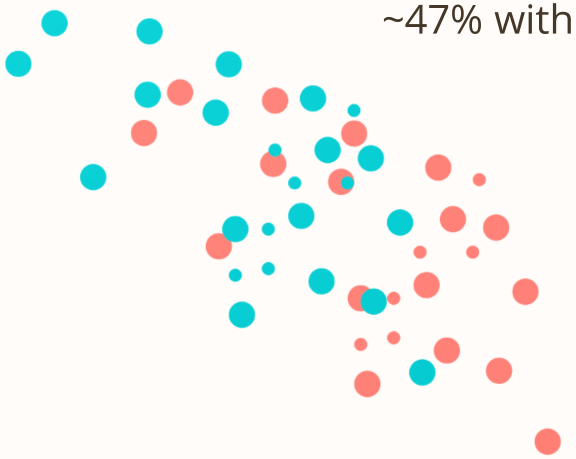




Target distribution shows a balanced dataset

~53% with heart disease,

~47% without.



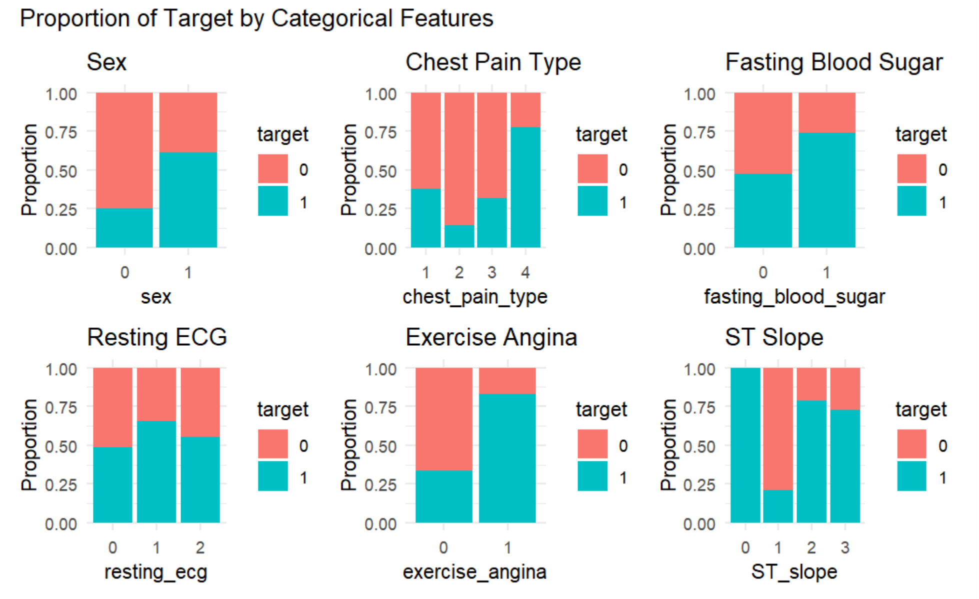
Distribution of Numeric Features

Variables like age, cholesterol, and max heart rate follow near-normal distributions.

Exploratory Data Analysis (EDA)

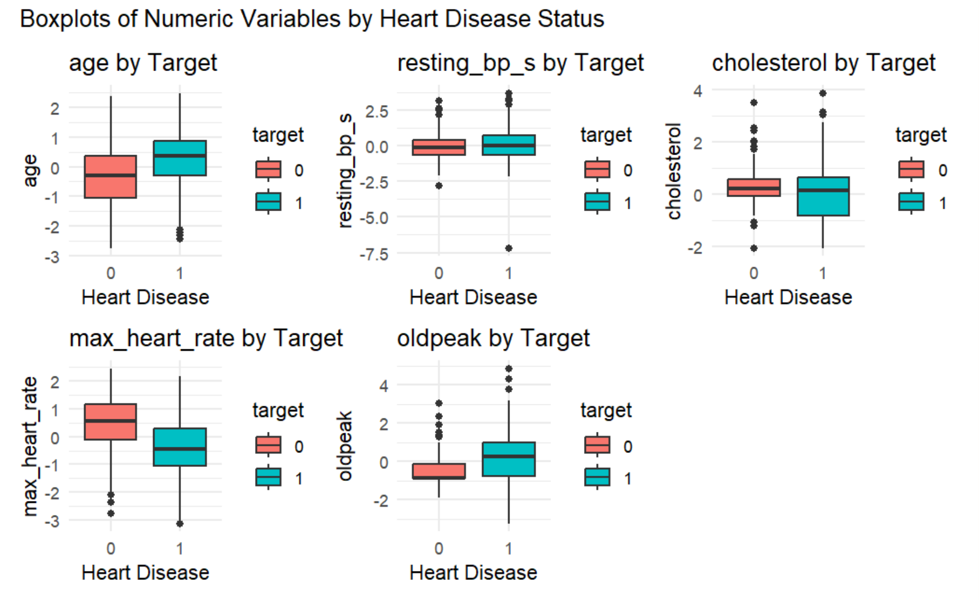
Proportion of Target By Categorical Features

Categorical features (e.g., chest pain type, exercise angina) show strong relationships with disease status.

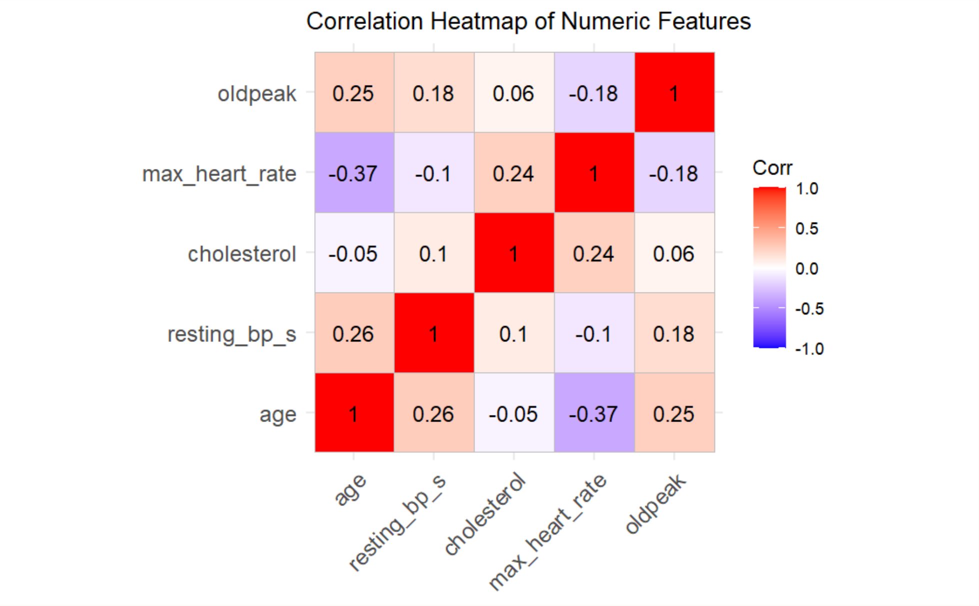


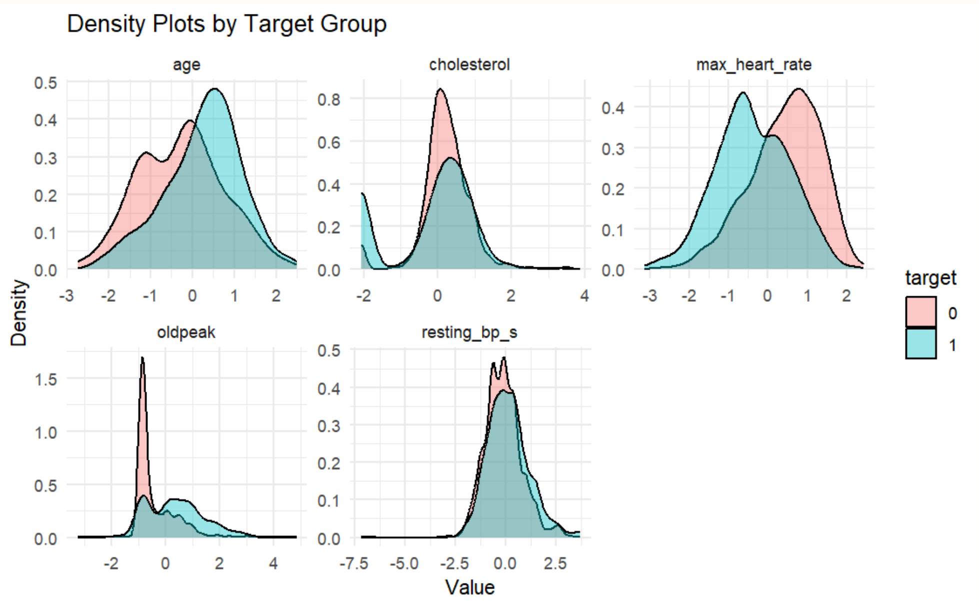
Boxplots of Numeric Variables by Heart Disease Status

Patients with heart disease generally have higher age and oldpeak, lower cholesterol and max heart rate than non-diseased individuals. Resting blood pressure shows overlap and may be less discriminative. Boxplots confirm variable separability and support model feature selection.



Exploratory Data Analysis (EDA)





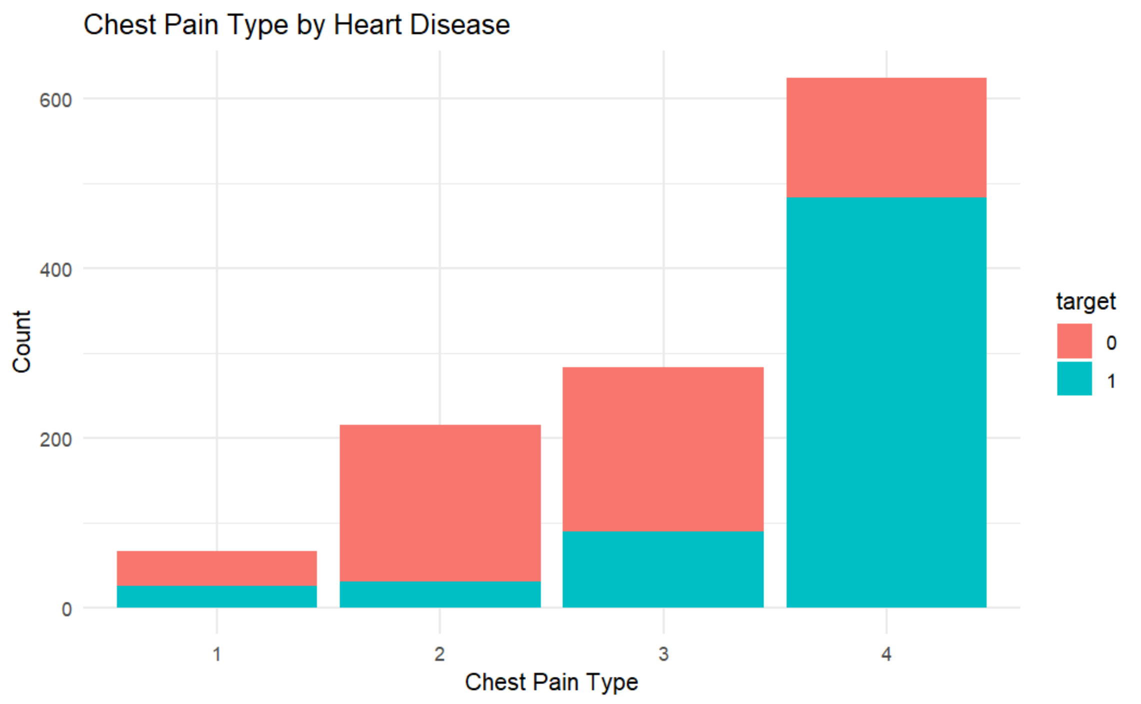
Correlation Heatmap of Numeric Features

Most variables exhibit low correlation (|r| < 0.4), suggesting minimal multicollinearity. Age and resting blood pressure show a mild correlation (~0.26). Max heart rate is negatively correlated with both age and oldpeak, indicating lower exertion capacity with age and ST depression.

Density Plots by Target Group

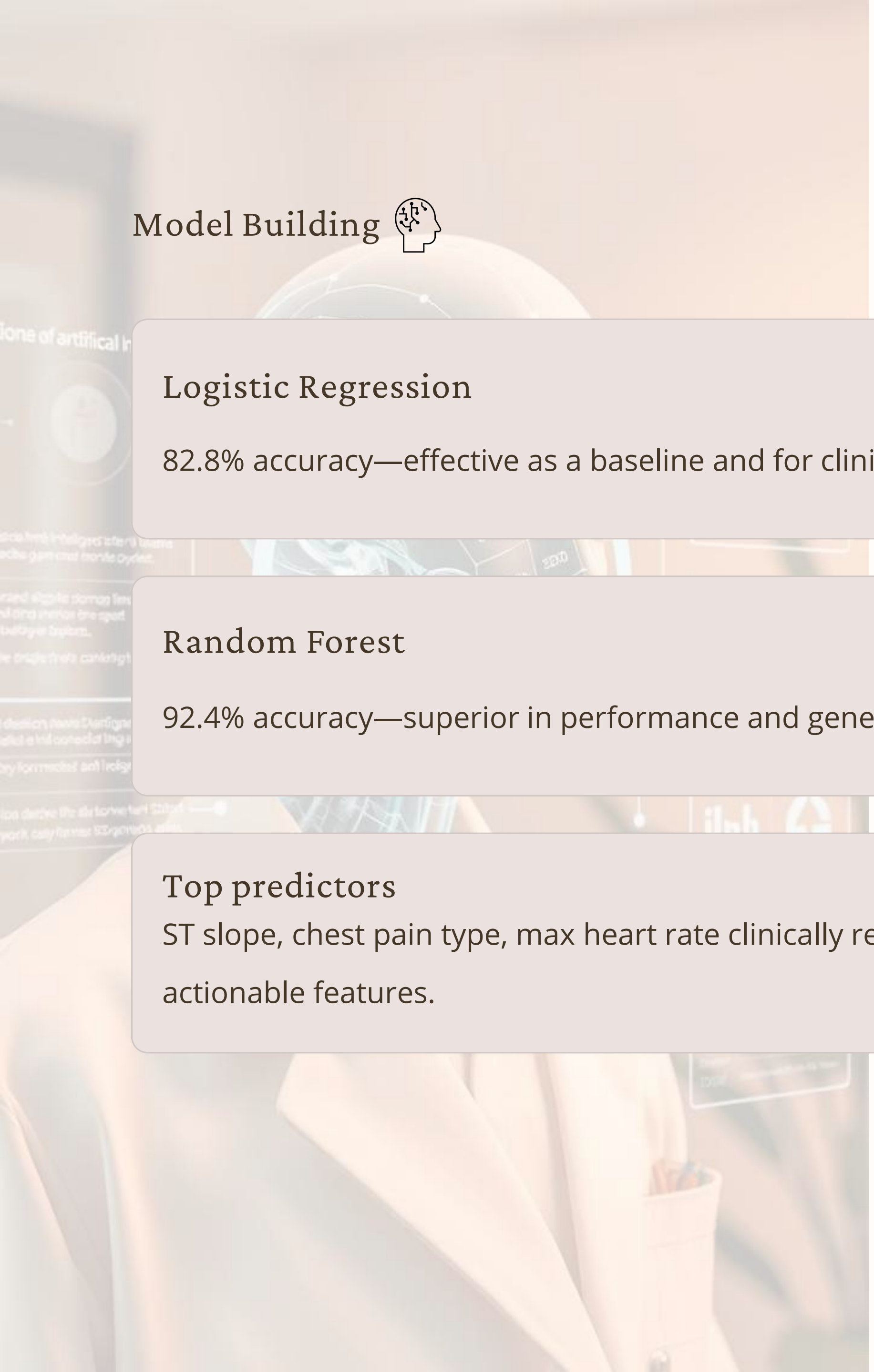
Heart disease patients tend to have:Higher age and oldpeak valuesLower max heart rate and cholesterolThese plots illustrate how feature distributions differ across target labels, reinforcing class separation.

Exploratory Data Analysis (EDA)



Chest Pain Type by Heart Disease Status

Chest pain type 4 (asymptomatic) is strongly associated with positive heart disease diagnoses. Type 3 also shows a significant portion of positive cases. A clear gradient of risk is observable across chest pain categories, indicating this feature's strong predictive power.



Model Building

Logistic Regression

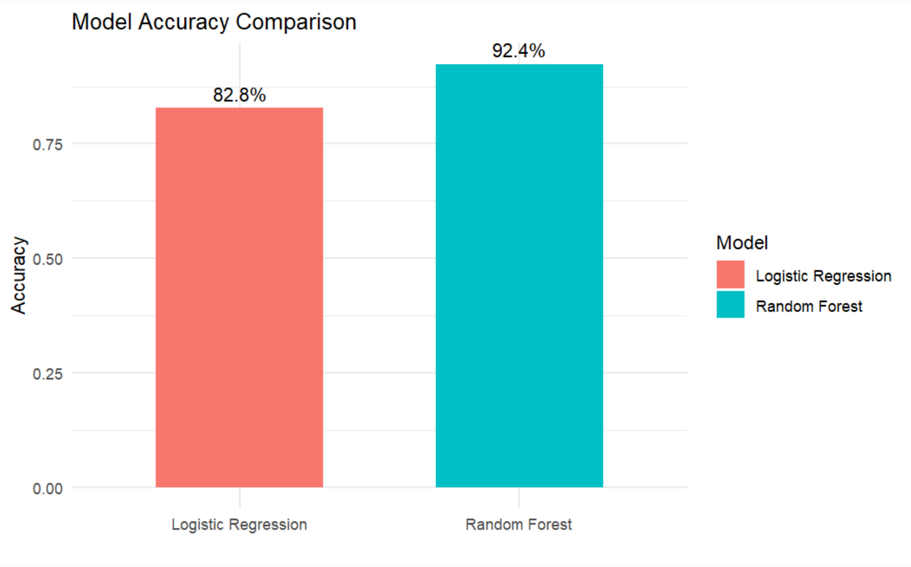
82.8% accuracy—effective as a baseline and for clinical explanation

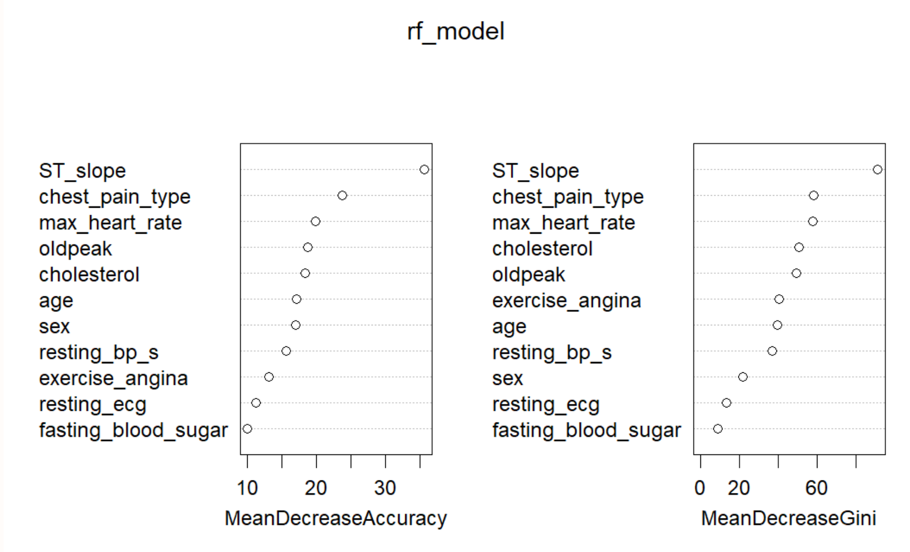
Random Forest

92.4% accuracy—superior in performance and generalization.

Top predictors

ST slope, chest pain type, max heart rate clinically relevant and actionable features.





Deployment, Ethics & Conclusion

Data Science Lifecycle (CRISP-DM Framework)

- Addresses late detection and diagnostic inefficiencies in CVD.
- Explores structured clinical datasets with numeric & categorical features.
- Performed encoding, normalization, and transformation.
- Applied Logistic Regression and Random Forest classifiers.
- Used accuracy, precision, recall, and cross-validation for model validation.
- Built a prototype Shiny app for real-time clinical decision support..

Clinical Readiness

- Interface: Shiny-based, user-friendly app for entering patient data and receiving predictions.
- Integration: HL7 FHIR compatible, suitable for EHR environments.
- Purpose: Designed for decision support, not standalone diagnosis.

Governance & Ethics

- Privacy & Security: Anonymized data; GDPR & ISO 27001 compliant setup.\
- Fairness: Bias analysis across gender, age, and clinical symptoms.
- Explainability: Incorporated SHAP & LIME for interpretable predictions.
- Compliance: Adheres to FAIR principles for data transparency and reuse.

Conclusion

- ML enables affordable, scalable, and early risk prediction for cardiovascular diseases.
- Random Forest demonstrates superior performance and clinical value.
- Future potential includes wearable integration and global-scale deployments to improve outcomes and optimize resources.