

HEART DISEASE PREDICTION

Nguyen Hoang Tung
Pham Cong Duyet

22BA13318
23BI14136

TABLE OF CONTENT

1

Introduction

2

Data Collection and
management

3

Visualization

4

Model Development

5

User Interface

6

Conclusion

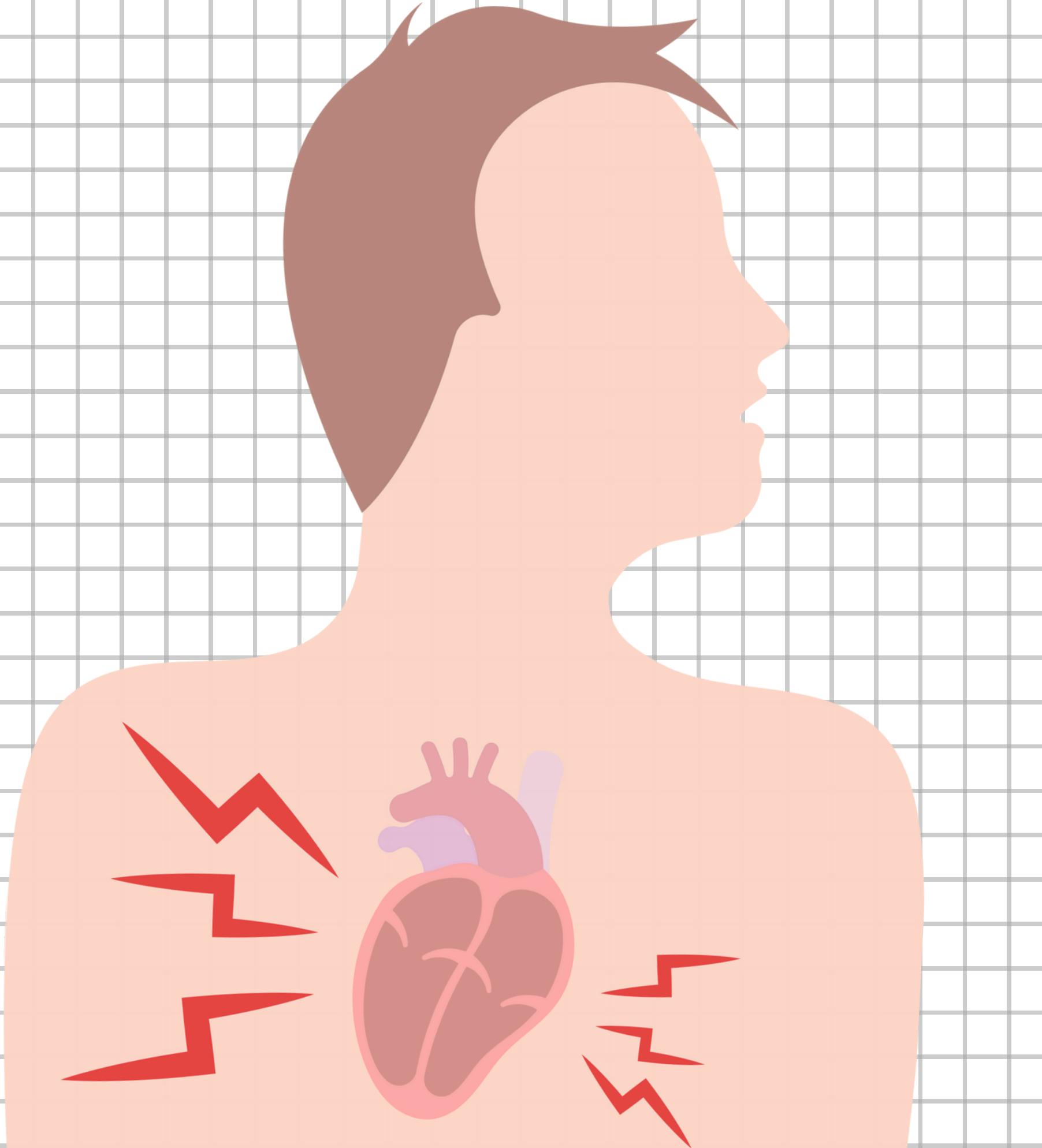
I. INTRODUCTION

Motivation

- Heart disease is one of the leading causes of death worldwide (~17.9 million deaths/year – WHO)
- In Viet Nam, 1 in 4 adults has high blood pressure.
- Heart disease ranks among the top causes of death and hospitalization.
- Many cases are detected late, leading to costly and less effective treatment

Business Problem

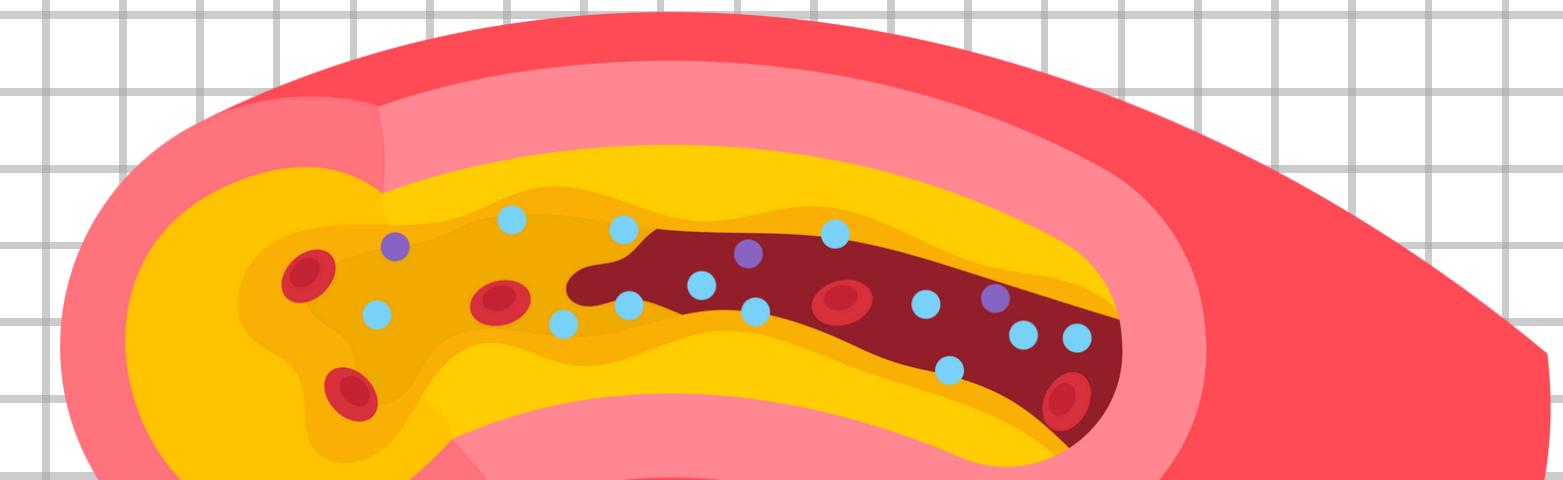
- **Risk Assessment and Prioritization:**
The model supports clinicians in assessing the probability of heart disease among examined patients, helping identify those at higher risk who may require closer monitoring, further diagnostic testing, or preventive treatment.
- **Cost and Resource Optimization:**
By highlighting high-risk individuals before their condition worsens, hospitals can allocate medical resources more efficiently, reduce unnecessary emergency admissions, and lower long-term treatment costs.
- **Evidence-Based Decision Support:**
The insights generated from predictive analytics assist healthcare managers and clinicians in making data-driven decisions, improving care quality, and supporting preventive strategies within hospital systems.



GOALS

- Collect and preprocess patient data.
- Explore and analyze the dataset to find patterns and risk factors.
- Train and evaluate multiple ML models to predict heart disease.
- Compare performance and select the best model.
- Build a simple Flask web interface for user-friendly predictions.

DATA COLLECTION AND MANAGEMENT



DATA COLLECTION AND MANAGEMENT

Data Collection

Dataset Overview

- Source: Kaggle – Heart Failure Prediction Dataset (Fedesariano, 2021)
- Samples: 918 records
- Features: 12 attributes (demographic + clinical)

Target Variable

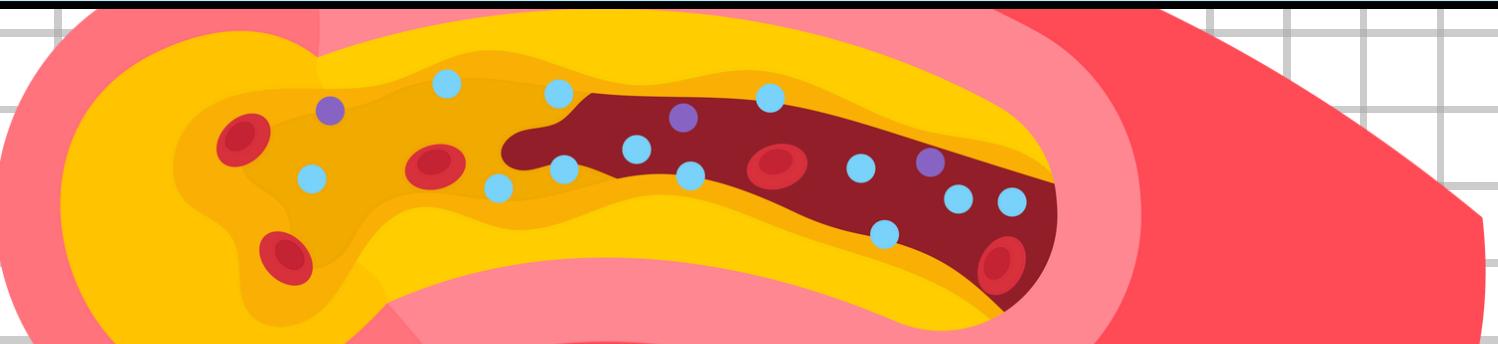
- HeartDisease is the output class used for prediction.
- 1 indicates disease presence; 0 indicates normal condition.

Key Features

- Age: patient's age (years)
- Sex: gender (M: Male, F: Female)
- ChestPainType: type of chest pain (TA, ATA, NAP, ASY)
- RestingBP: resting blood pressure (mmHg)
- Cholesterol: serum cholesterol level (mg/dl)
- FastingBS: fasting blood sugar (1 if >120 mg/dl, else 0)
- RestingECG: resting ECG result (Normal, ST, LVH)
- MaxHR: maximum heart rate achieved (60–202 bpm)
- ExerciseAngina: exercise-induced angina (Y/N)
- Oldpeak: ST depression induced by exercise (numeric)
- ST_Slope: slope of peak exercise ST segment (Up, Flat, Down)
- HeartDisease: target variable (1 = heart disease, 0 = normal)

DATASET

Age	Sex	CPT	R.BP	Chol	F.BS	R.ECG	MaxHR	EA	Oldpeak	Slope	HS
40	M	ATA	140	289	0	NORMAL	172	N	0	UP	0
49	f	NAP	160	180	0	NORMAL	156	N	1	FLLAT	1
37	M	ATA	130	283	0	ST	98	N	0	UP	0
48	F	ASY	138	214	0	NORMAL	108	Y	1.5	FLAT	1



DATA COLLECTION AND MANAGEMENT

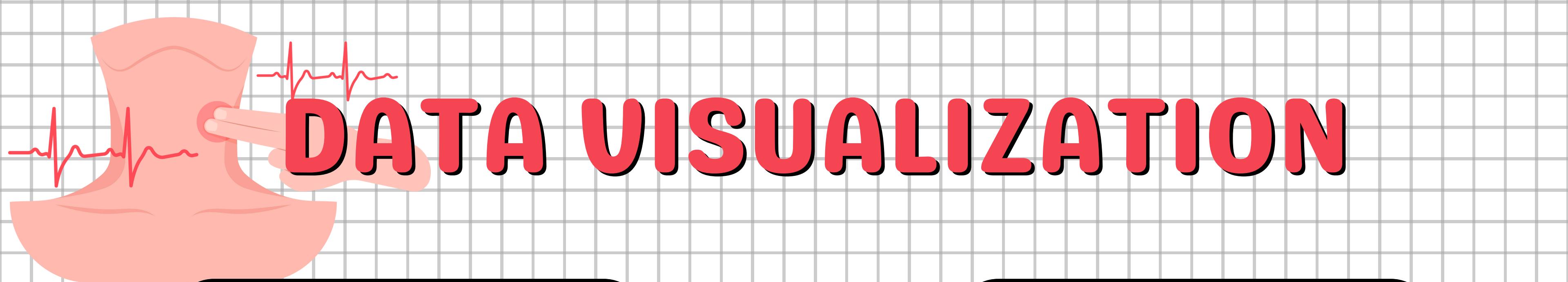
Data management

Data Cleaning

- Identified invalid or unrealistic values in the dataset.
- RestingBP: one record = 0 → replaced with median (since BP can't be 0).
- Cholesterol: 172 records = 0 → replaced using KNN Imputer to estimate realistic values from similar patients.
- Other features (Age, Sex, ChestPainType, RestingECG, etc.) were clean and consistent.
- This step ensured data reliability and reduced bias during model training.

Data Processing

- Handled categorical variables that machine learning models can't process directly.
 - Binary Encoding:
 - Sex: M→1, F→0
 - ExerciseAngina: Y→1, N→0
 - One-Hot Encoding:
 - ChestPainType, RestingECG, ST_Slope → converted to multiple dummy columns.
- Result: dataset transformed into fully numerical and standardized format for modeling. It consists of 19 columns including the target column.



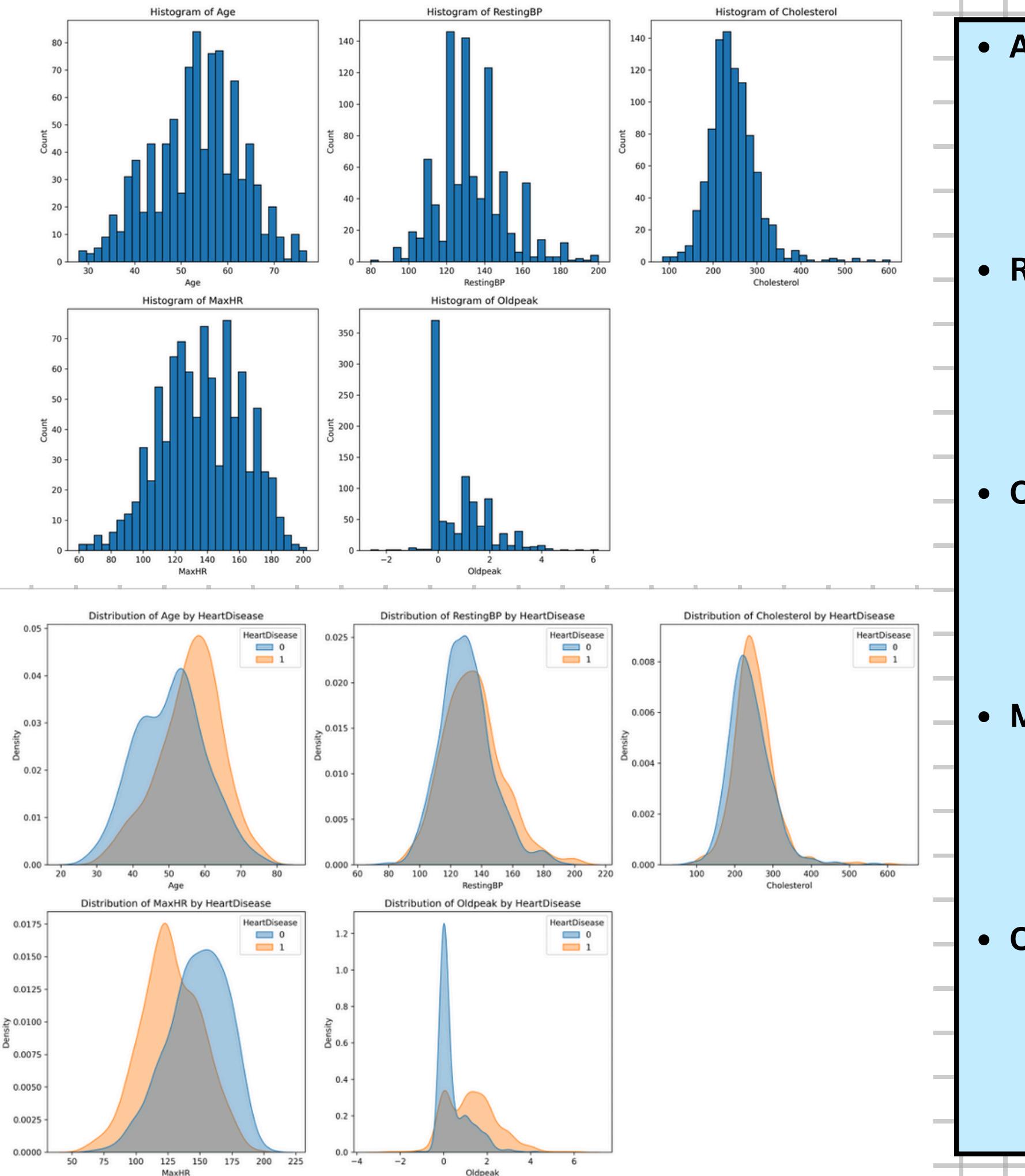
DATA VISUALIZATION

Visualization Goals

- Explore the data distribution of key variables
- Compare categorical features
- Identify patterns and potential predictors of heart disease

Visualization Methods

- Histograms
- KDE plots
- Countplots
- Boxplots
- Correlation heatmap



• Age

- Distribution: Concentrated between 45–65 years, with diseased patients peaking around 55–65.
- Insight: Middle-aged and older adults have higher exposure to risk factors → higher heart disease prevalence.

• RestingBP

- Distribution: Most values around 120–140 mmHg, similar across both groups.
- Insight: Hypertension is common but not highly discriminative in this dataset.

• Cholesterol

- Distribution: Left-skewed; most below 300 mg/dl, few extreme cases >400.
- Insight: Overlaps between groups; may be affected by medication or lifestyle, thus limited predictive power.

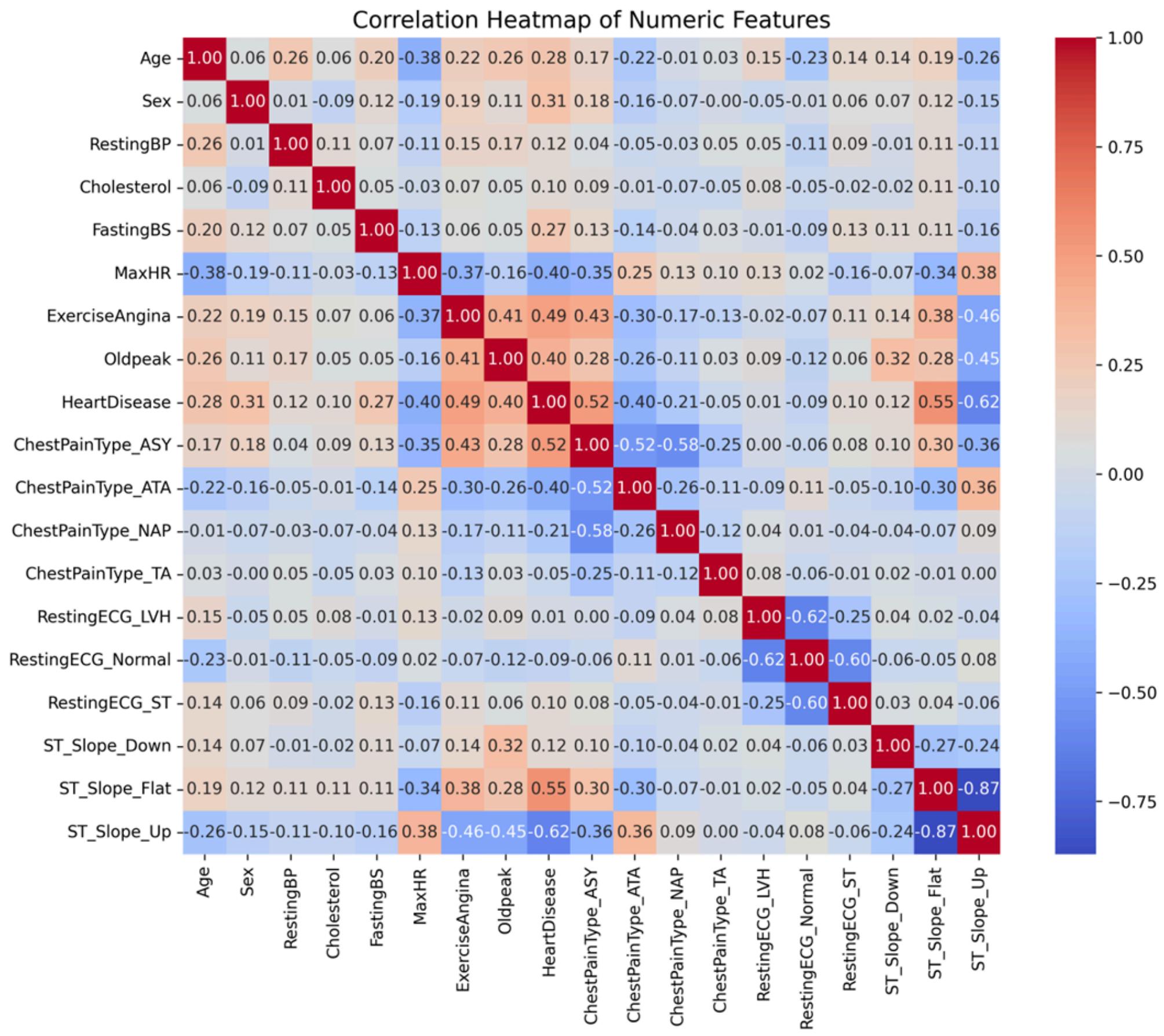
• MaxHR

- Distribution: Approximately normal, centered 120–160 bpm; diseased patients cluster below 140 bpm.
- Insight: Lower MaxHR reflects reduced exercise tolerance – a strong clinical marker of coronary disease.

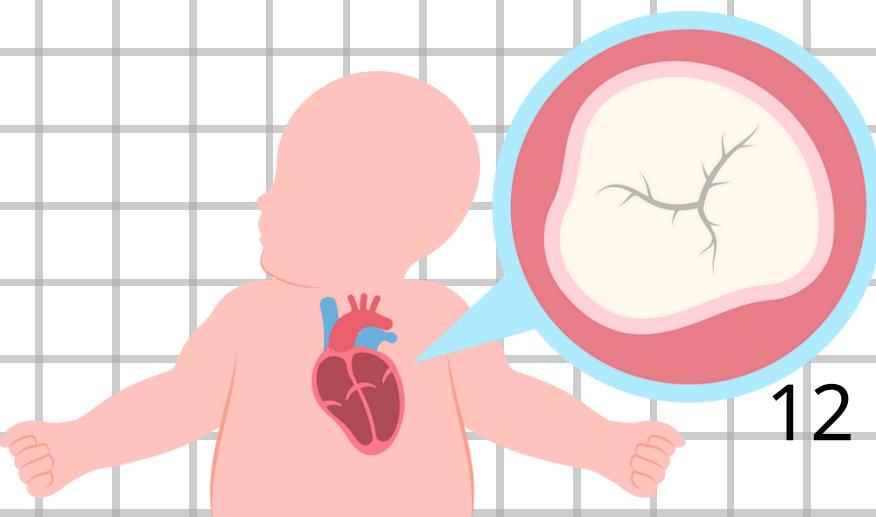
• Oldpeak

- Distribution: Most healthy patients ≈ 0; diseased patients show higher values (up to >4).
- Insight: ST depression (Oldpeak) strongly indicates myocardial ischemia – one of the most powerful predictors.

- Oldpeak, ST_Slope, and ExerciseAngina are the strongest predictors of heart disease.
- MaxHR and Age show moderate effects, while RestingBP, Cholesterol, and FastingBS have weak or no correlation, likely due to treatment or sample limits.

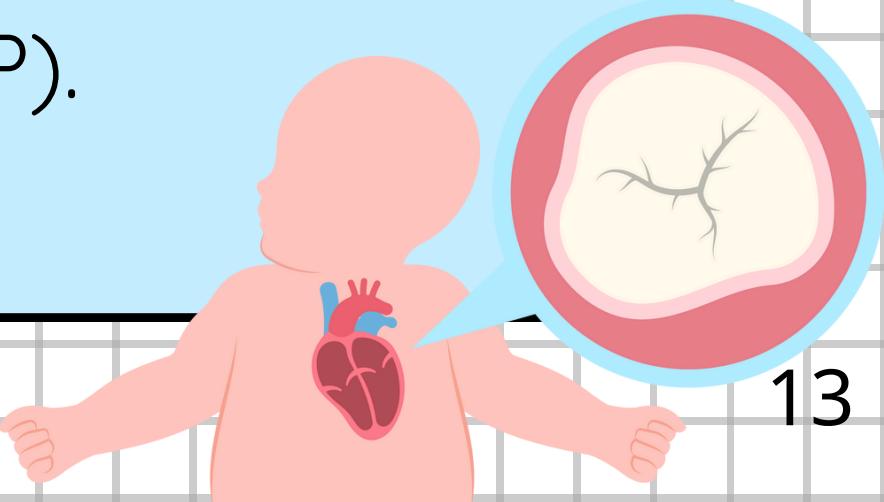


MODEL DEVELOPMENT



Data Preparation

- Dataset divided 80/20, stratified by HeartDisease to maintain class balance and ensure reproducibility.
- Scale continuous features with **StandardScaler**.
- Automates preprocessing before model fitting.
- Prevents leakage and supports model flexibility (Logistic Regression, SVM, RF, XGBoost).
- Cross-Validation:
- Stratified K-Fold ($k=5$) ensures balanced label distribution → robust evaluation.
- Transformed features re-combined into DataFrame →
- enables post-hoc analysis (e.g., Permutation Importance, SHAP).



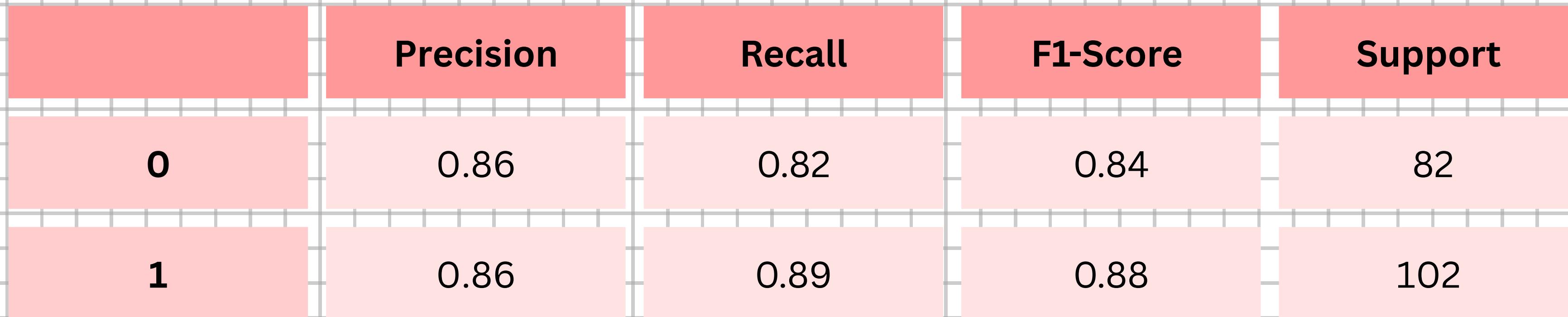
LOGISTIC REGRESSION

	Precision	Recall	F1-Score	Support
0	0.85	0.88	0.86	82
1	0.90	0.87	0.89	102

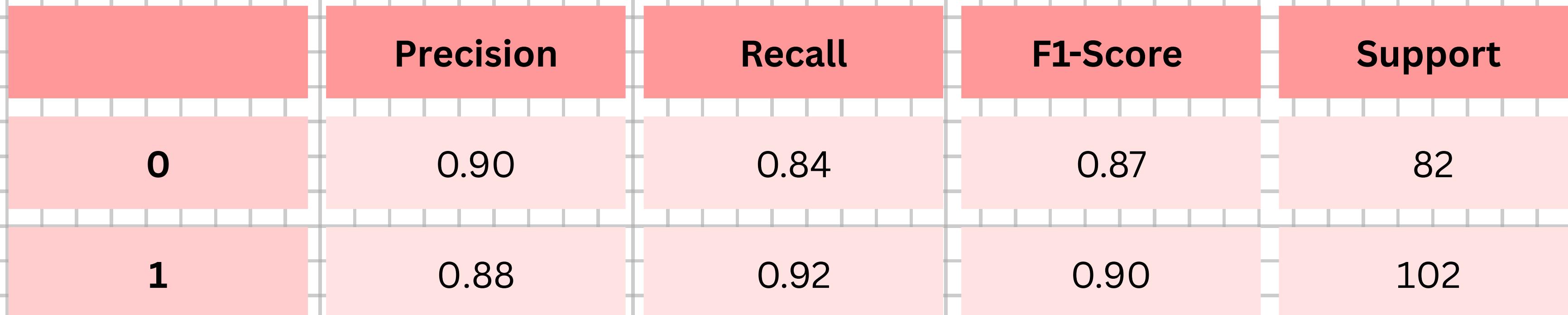
SVC

	Precision	Recall	F1-Score	Support
0	0.83	0.87	0.85	82
1	0.89	0.85	0.87	102

RANDOM FOREST



XGB



ACCURACY

Model	CV ROC-AUC	TEST ROC-AUC	Accuracy
LR	0.9221	0.9320	0.88
SVC	0.9201	0.9366	0.86
RF	0.9290	0.9229	0.86
XGB	0.9328	0.9318	0.89

SUMMARY OF RESULT

Model	CV ROC-AUC	TEST ROC-AUC	Accuracy
LR	0.9221	0.9320	0.88
SVC	0.9201	0.9366	0.86
RF	0.9290	0.9229	0.86
XGB	0.9328	0.9318	0.89

Logistic Regression (LR)

- Achieved the high accuracy (0.88) and a strong ROC-AUC (0.9320).
- Results are highly consistent between CV and test sets, indicating stable generalization.

Support Vector Machine (SVC)

- Reached the highest Test ROC-AUC (0.9366), showing superior discriminative ability.
- However, accuracy (0.86) is slightly lower, and performance variation between CV and test sets suggests less stability.

Random Forest (RF)

- Produced a solid CV ROC-AUC (0.9290) and comparable Test ROC-AUC (0.9229), showing acceptable generalization.
- Minor performance gap indicates slight overfitting tendency, but still performs reliably across datasets.

XGBoost (XGB)

- Demonstrated the most balanced and consistent performance overall.
- Achieved the highest CV ROC-AUC (0.9328) and a strong Test ROC-AUC (0.9318) with accuracy = 0.89.
- Results are stable across folds and test sets, proving excellent robustness and generalization.

🎯 Conclusion

- Logistic Regression → Best for simplicity, interpretability, and solid accuracy.
- SVC → Strongest discriminative ability, ideal when precision is prioritized.
- Random Forest → Reliable ensemble with mild overfitting risk.
- XGBoost → Best trade-off between accuracy, stability, and robustness → recommended final model.

◆ Top Predictive Features

ST_Slope (Up) and ChestPainType_ASY show the strongest impact on model performance.

→ When shuffled, they cause the largest ROC-AUC drop

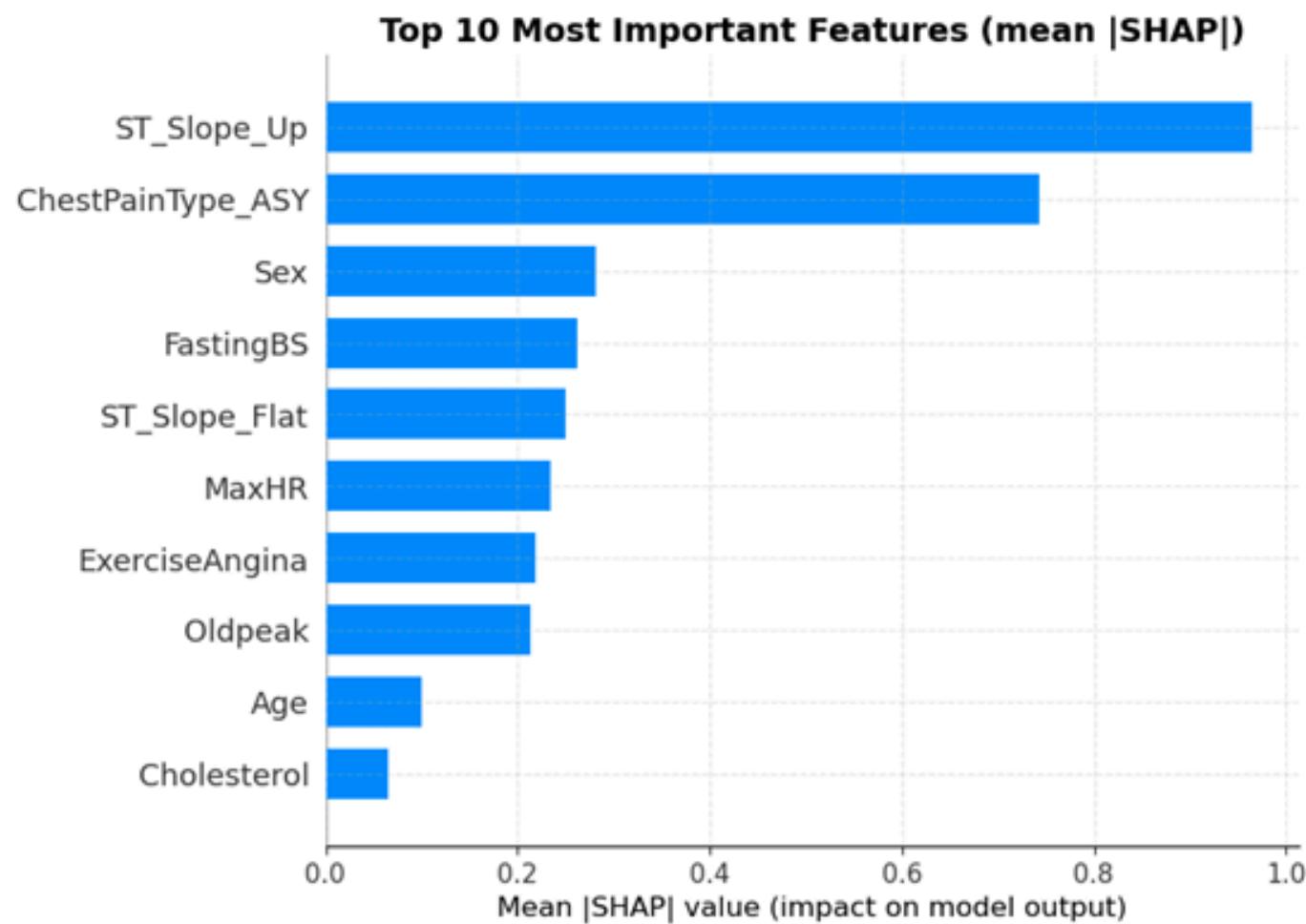
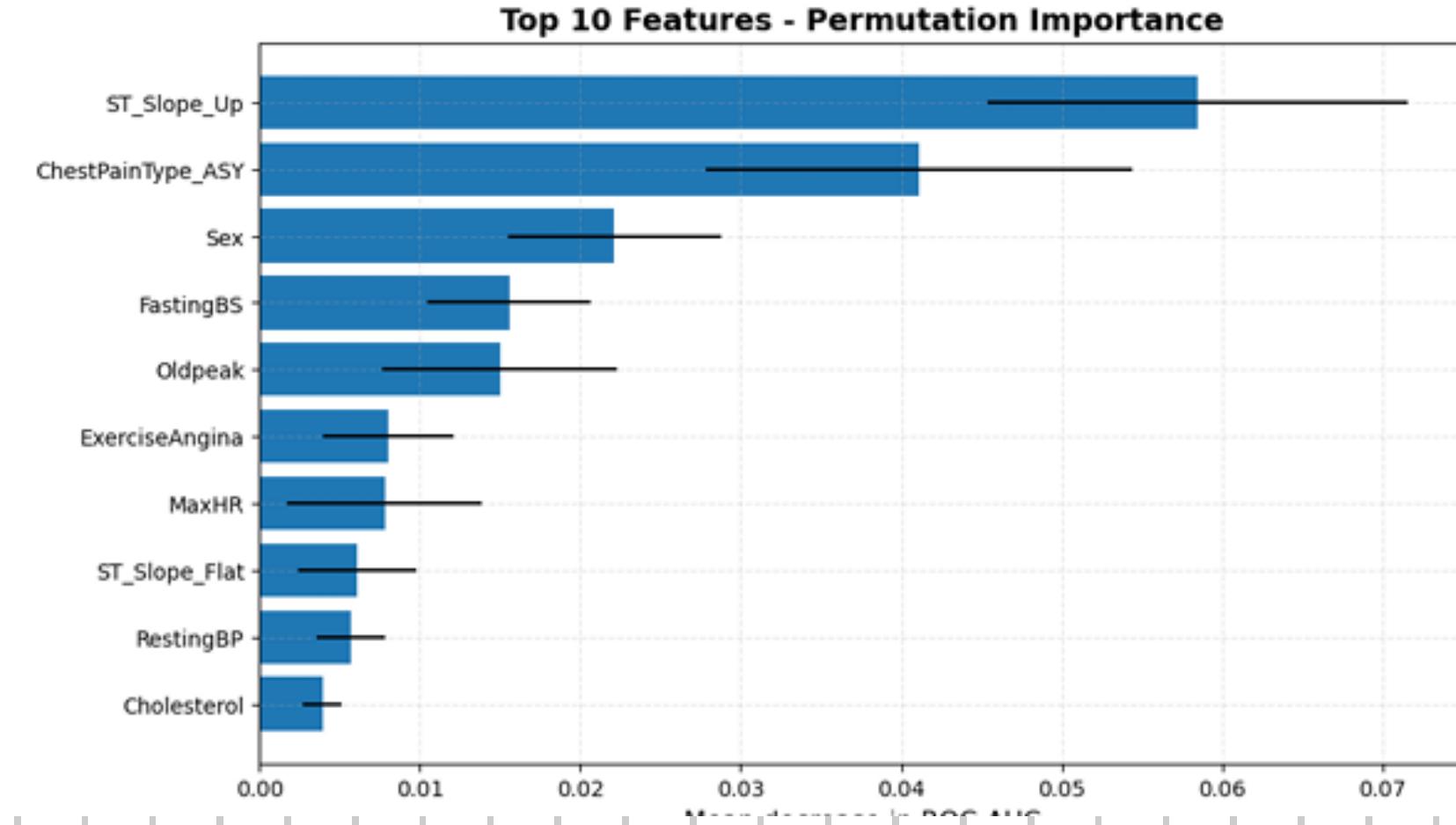
◆ Moderate Predictors

Sex and FastingBS show moderate contributions, suggesting that biological and metabolic differences still play a role but are secondary to ECG-based measures.

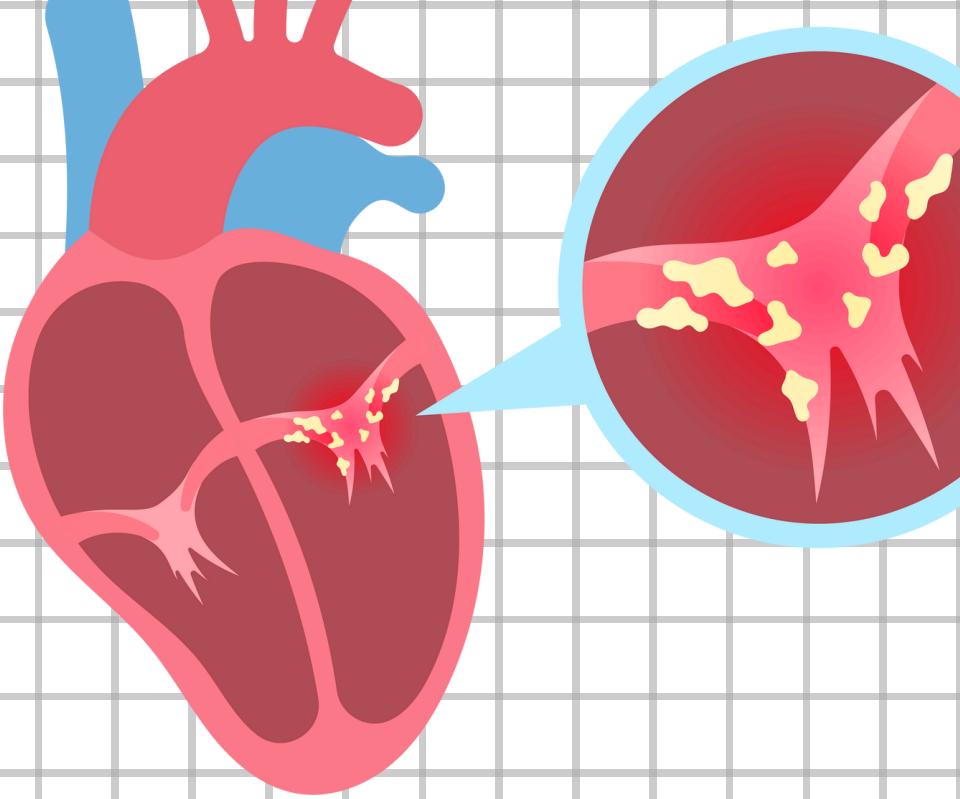
◆ Low-Contributing Features

Age, Cholesterol, and RestingBP contribute minimally to prediction accuracy.

→ This may be due to dataset composition (controlled or medicated patients) or the model's tendency to rely on dynamic physiological signals rather than static clinical measurements.



USER INTERFACE



app.py

- The trained XGBoost pipeline was exported as xgb_pipeline.pkl using Joblib (faster and more efficient than Pickle).
- A Flask-based web app was developed for heart disease prediction.
- Users input clinical data (age, sex, blood pressure, cholesterol, ECG results, etc.).
- Results are displayed on the web page (index.html).
- If invalid data is entered, the app shows an error message instead of breaking.

index.html

- Developed using HTML and Bootstrap 5 for simplicity, responsiveness, and modern UI styling.
- Blue gradient background, white rounded form with shadow, and blue buttons for a clean, medical-themed aesthetic.
- Contains 11 main fields
- Includes tooltips to explain each field's meaning and assist user input.
- User Experience Enhancements:
 - Displays a loading spinner while processing predictions.
 - Results (probability + Positive/Negative label) appear in a dedicated result box below the form.

CONCLUSION

1. Project Overview

- Objective: Develop a machine learning-based system for predicting the probability of heart disease.
- Core contribution: Integrating a robust classification model (XGBoost) with a user-friendly web interface, demonstrating the potential of ML in healthcare.
- Results:
 - → High model performance on structured data.
 - → Real-time prediction via Flask web app enhances accessibility.

2. System Implementation

- Backend: Flask application connected to the XGBoost model saved via Joblib for fast loading.
- Frontend: Built with HTML and Bootstrap, featuring tooltips, validation, and a loading spinner.
- Process: User inputs → validation → transformation → prediction (Positive/Negative + probability).
- Demonstrates how ML can support real-time, interpretable healthcare decision-making.

3. Key Insights

- The model performs stably and generalizes well to unseen data.
- The interface design ensures ease of use, even for non-technical users.
- The project proves the practical value of predictive analytics in modern healthcare

4. Future Work

1. Data Expansion: Add more samples and clinical variables to improve model robustness.
2. Explainability (SHAP): Integrate model explainability to help clinicians understand prediction logic.
3. Cloud Deployment: Move the system to a cloud environment for scalability and accessibility.

5. Final Remark

This project successfully built a reliable and interpretable heart disease prediction system while laying the foundation for further integration of machine learning, web development, and healthcare analytics – bridging the gap between data science and clinical decision support.



GITHUB

<https://github.com/tungtimo0808/Heart-Failure-Prediction>

