Data Dictionary

Feature Name	Description	Type	Example
Age	Age of the patient	Numerical	54
Sex	Gender of the patient (0 = Female, 1 = Male)	Categorical	1
ChestPainType	Type of chest pain (TA, ATA, NAP, ASY)	Categorical	NAP
RestingBP	Resting blood pressure (mm Hg)	Numerical	130
Cholesterol	Serum cholesterol (mg/dL)	Numerical	246
FastingBS	Fasting blood sugar > 120 mg/dL (0 = No, 1 = Yes)	Categorical	0
RestingECG	ECG results (Normal, ST, LVH)	Categorical	Normal
MaxHR	Maximum heart rate achieved	Numerical	150
ExerciseAngina	Exercise-induced angina (0 = No, 1 = Yes)	Categorical	0
Oldpeak	ST depression induced by exercise	Numerical	1.4
ST_Slope	Slope of the peak exercise ST segment (Up, Flat, Down)	Categorical	Flat
HeartDisease	Target variable (0 = No, 1 = Yes)	Binary Target	1

Preprocessing Steps

1. Missing/Invalid Value Handling:
    - Removed or replaced entries with invalid values (e.g., RestingBP or Cholesterol = 0).
2. Encoding Categorical Features:
    - Label Encoding for binary features: Sex, ExerciseAngina
    - One-Hot Encoding for multi-class features: ChestPainType, RestingECG, ST_Slope
3. Feature Scaling:
    - Applied StandardScaler to numerical features (e.g., Age, Cholesterol, RestingBP, MaxHR, Oldpeak) to normalize input ranges.
4. Splitting Features/Target:
    - Dataset split into X (features) and y (HeartDisease target)
5. Train-Test Split:
    - Split data into training and testing sets (e.g., 80/20 split) for model evaluation**.**

Model Explanation

Several machine learning models were trained and evaluated:

- Logistic Regression:
  - A linear model used for binary classification.
  - Interpretable and fast, but may underperform on complex relationships.
- Random Forest Classifier:
  - An ensemble of decision trees using bagging.
  - Robust, handles both numerical and categorical data well.
- XGBoost:
  - Gradient Boosted Trees optimized for performance.
  - Excellent for structured/tabular data and imbalanced problems.
- Neural Network (Keras):
  - Multi-layer perceptron used to capture complex patterns.
  - Requires careful tuning and normalization.

✅ Final selection depends on evaluation metrics like accuracy, precision, recall, and F1-score.

User Guide

Overview:

The app allows users to input patient data and get a real-time prediction on heart disease risk.

Pages Overview:

1. Home Page:
   - Introduces the app and explains its purpose.
2. Preprocessing Page:
   - Shows how raw data is cleaned and transformed.
3. Visualizations Page:
   - Displays EDA charts (distribution, correlations) for insight into data behavior.
4. Modeling Page:
   - Allows user input of medical parameters
   - User selects model (e.g., Random Forest, XGBoost)
   - App outputs prediction and probability instantly

How to Use:

- Navigate to the Modeling Page
- Enter values for all required medical fields
- Select a model from the dropdown
- Click "Predict"
- View result:
    - ✅ No heart disease
    - 🔴 Likely heart disease (with probability %)

System Requirements:

- Just a web browser
- No installation needed
- Hosted via Streamlit Cloud