

# Heart Disease Prediction Dataset — Analytical Context Document

## 1. Dataset Overview

This dataset contains **270 patient records** with **14 clinical attributes** related to cardiovascular health.

Each record represents a single patient encounter. The dataset is designed to support **exploratory analysis and predictive modeling** of heart disease outcomes.

The target variable, **Heart Disease**, indicates whether heart disease is **Present** or **Absent**.

---

## 2. Feature Description

The dataset includes demographic, clinical, and exercise-related indicators:

- **Age** – Patient age in years
  - **Sex** – Gender indicator (binary encoded)
  - **Chest pain type** – Categorized chest pain levels
  - **BP** – Resting blood pressure
  - **Cholesterol** – Serum cholesterol level
  - **FBS over 120** – Fasting blood sugar > 120 mg/dl (binary)
  - **EKG results** – Resting electrocardiographic results
  - **Max HR** – Maximum heart rate achieved during exercise
  - **Exercise angina** – Exercise-induced angina (binary)
  - **ST depression** – ST depression induced by exercise
  - **Slope of ST** – Slope of the peak exercise ST segment
  - **Number of vessels fluro** – Number of major vessels colored by fluoroscopy
  - **Thallium** – Thallium stress test result
  - **Heart Disease** – Target outcome (Presence / Absence)
-

### 3. Analytical Objective

The primary objective is to **analyze relationships between clinical indicators and heart disease outcomes**, with a focus on:

- Understanding how **exercise-related features** behave across outcome groups
- Identifying **patterns, distributions, and co-occurring features**
- Supporting **binary classification** and **risk-pattern exploration**

This dataset is **not intended to establish clinical causality**, but rather to support **data-driven pattern analysis**.

---

### 4. Exercise-Related Indicators (Key Focus Area)

Several features capture patient response to exercise stress:

- Exercise angina
- ST depression
- Slope of ST
- Maximum heart rate

These indicators are commonly analyzed together to understand **exercise tolerance and stress response** in relation to heart disease outcomes.

---

### 5. Interpretation Guidelines (Critical)

To avoid misinterpretation:

- **Raw frequency or count of feature combinations reflects data distribution, not risk**
- Associations must be evaluated by **comparing outcome groups (Presence vs Absence)**

- Rare feature combinations should not be interpreted as protective or harmful without normalization
  - Observations describe **patterns**, not medical conclusions
- 

## 6. Example Analytical Questions

This dataset supports questions such as:

- How do exercise-related indicators differ between patients with and without heart disease?
- Which clinical features tend to co-occur more frequently in heart disease cases?
- Are there distinct exercise-response patterns associated with heart disease outcomes?

These questions require **multi-feature reasoning** rather than single-column lookup.

---

## 7. Intended Use in Analytical Systems

This dataset is suitable for:

- Exploratory Data Analysis (EDA)
- Feature interaction analysis
- Binary classification modeling
- Evaluation of retrieval-augmented analytical systems