



**DEPARTMENT OF APPLIED ARTIFICIAL  
INTELLIGENCE**

**SHILEY-MARCOS SCHOOL OF ENGINEERING**

**UNIVERSITY OF SAN DIEGO, SAN DIEGO,  
CALIFORNIA**

**UNITED STATES OF AMERICA**

**2025**

# **AI-Based Heart Disease Prediction Using Patient Reports**

A Project Report

*Submitted in partial completion of the requirements for the award of  
the degree of*

**MASTER OF SCIENCE  
IN  
APPLIED ARTIFICIAL INTELLIGENCE**

**Submitted by:**

***Sharon Karabel***

**Under the Supervision of:**

***Dr. Azka Azka***

# Abstract

**Keywords:** Predictive Analytics, Clinical Decision Support, Health Informatics, Risk Stratification, Model Explainability.

Cardiovascular disease (CVD) remains the top contributor of death globally, highlighting the urgent need for precise, data-driven risk prediction systems. This research aims to implement AI methods to estimate the possibility of heart diseases using the structured clinical dataset Heart.csv, which includes information of the patients namely age\_group, gender, lipid profile, pressure of blood flow, and electrocardiographic findings.

I developed and evaluated multiple predictive systems, including Random\_Forest, Naïve-Bayes, & ensemble techniques, specifically XGBoost. These models were tuned to categorize patient risk levels and identify patterns associated with cardiovascular events. The random\_forest achieves the highest precision and robustness across validation sets.

To ensure clinical relevance and trust, we incorporated model explainability tools such as LIME values and feature importance analysis, which consistently highlighted features like ST\_segment, MaxHR, Oldpeak, and resting ECG results as top contributors to predictions.

Additionally, a user-centric web interface was created to allow medical practitioners to add individual-specific data and obtain real-time, individualized risk assessments. This approach demonstrates the potential of intelligent health systems to support proactive cardiovascular support and facilitate timely clinical interventions.

## Acknowledgment

In acknowledgment of this project's achievement, numerous individuals have offered me their blessings and heartfelt support.

Firstly, I want to convey my thanks to the Almighty for allowing me to submit this research successfully. I am sincerely thankful to my instructor for her encouragement and help, **Dr. Azka Azka**, whose insightful guidance has been paramount in helping me finish this project and achieve complete success. Her recommendations and guidance have been the primary factor in the project's development.

Ultimately, I recognize the wider scientific community whose continuous efforts in healthcare analytics and AI keep motivating and aiding innovation to enhance patient outcomes.

## Table of Contents

Section No.	Title	Pagination
1	Abstract	3
2	Acknowledgment	4
3	Table of Contents	5
4	Introduction	6
5	Problem Definition	6
6	Objective	7
7	Dataset Overview	8
8	Libraries used	9
9	Data Pre-processing and cleaning	10
10	Statistical Description	12
11	Exploratory Data Analysis	15
12	Feature Engineering for EDA	34
13	Feature Importance Analysis	36
14	Hypothesis Testing	37
15	Contingency Table	41
16	Anomaly Detection	43
17	Construction for Predictive Models	47
18	Application of SMOTE	51
19	Evaluation of models	53
20	Model Deployment-Interactive Dashboard	56
21	Limitations	57
22	Future Works	58
23	Conclusions	59
24	Reference	60

#### **4. Introduction:**

This project focuses on developing an AI-based system to predict how heart disease patients will respond to various treatments using demographic and clinical data. By applying AI methods, the system aims to provide personalized predictions to assist healthcare providers in making knowledgeable choices, eventually enhancing patient results & enhancing the quality of HD management.

The system analyzes patterns in patient-specific information to identify key factors influencing treatment effectiveness. This enables proactive care planning, reduces the risk of adverse events, and supports precision medicine. The integration of AI in clinical workflows promises to revolutionize cardiovascular care by making it more tailored, data-driven, and outcome-focused.

#### **5. Problem Definition:**

Heart disease continues to be a notable communal health concern & is the primary contributor to global death rates. Initial diagnosis of heart-related conditions is essential to improve patient reports, minimizing hospitalization rates, and minimizing the long-term burden on healthcare systems. However, identifying at-risk individuals based on diverse clinical indicators can be complex and error-prone without computational support.

This project addresses the need for a reliable, interpretable, and accessible prediction system for heart failure using ML techniques. Utilizing the Heart\_Failure\_Prediction dataset, which includes demographic and clinical attributes like age\_group, biological identity, lipid levels, type of chest-pain, & Electro-cardiogram results, we implemented and compared several supervised learning classifiers like Random\_Forest, Naive-Bayes, & LR to find the existence of HD.

To improve the system's transparency & foster medical trust, we applied LIME i.e, Locally interpretable technique for model explanations, to clarify separate forecasts and highlight influential features contributing to model decisions. Additionally, we developed an intuitive web-based interface that allows medical professionals to input the information of the patient and obtain immediate risk predictions along with interpretable insights.

The predictive system will enable healthcare professionals to tailor treatment plans based on patients' predicted responses, thereby optimizing therapeutic outcomes and minimizing potential side effects. The model's deployment in clinical settings will facilitate personalized medicine approaches, enhancing patient care and quality of life. This approach seeks to assist clinicians in making informed, data-driven choices, thus promoting early intervention and improved management of cardiovascular illness.

Our goal involves developing an AI model to predict heart disease patients' responses to treatment based on demographic, genetic, and clinical data. The goal is to personalize care by identifying the most effective strategies for every individual, improving outcomes, and reducing risks. By analyzing patient-specific information-medical background, genetic markers, & clinical indicators—the model aims to support healthcare providers in making knowledgeable choices, resulting in improved patient treatment in heart disease management.

## 6. Objectives:

The goal of the project is to

- **To build ML models** which precisely forecasts the probability of heart disease using organized medical information.
- **To identify and analyse significant threat agents which contributed to HD** by evaluating feature importance and statistical associations.
- **To support medical decisions** by providing an AI-based diagnostic tool that enhances early detection and personalized intervention strategies.
- **To utilize data visualization techniques** for uncovering hidden patterns, trends, and correlations among patient health indicators.
- **To implement anomaly detection algorithms** for recognizing atypical patient profiles or rare cases that might represent unique clinical insights or data errors.

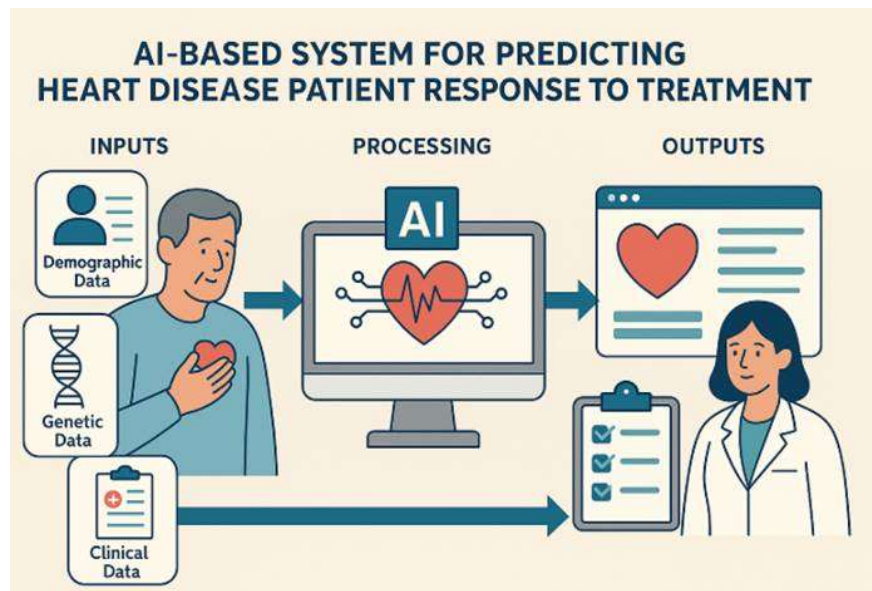


Fig (6.1)

## 7. Dataset Overview:

The Heart Failure Prediction dataset is commonly utilized in clinical data science to assess the chances of Heart\_Disease by examining a mix of demographic and physiological characteristics. In this project, we utilize a polished version called heart.csv, containing 918 patient entries with pertinent medical characteristics. The dataset originates from Kaggle and has been processed and normalized for application in machine learning classification projects. Our goal is to utilize artificial intelligence methods to help in the initial diagnosis of cardiovascular diseases, enabling healthcare practitioners to identify at-risk patients through organized data. The dataset contains categorical and numerical variables, including age, gender, lipid level, type of chest-pain, & ECG outcomes—vital factors in clinical decision-making.

This real-world data-set was constructed by combining 5 previously independent Heart\_Disease data samples that had not been merged before. The resulting dataset integrates 11 common features across all sources, which makes the biggest dataset exist at present for research purposes. These 5 data samples curated for this compilation are:

Location	Number of Records
Cleveland	303
Hungary	294
Switzerland	123
Long Beach VA	200
Stalog (Heart) Data Set	270

**Overall:** 1190 records

**Duplicated:** 272 records

**Result:** 918 records



## 8. Libraries Utilized:

We have imported the following libraries for our project,

- a. **pandas & NumPy**: Used for efficient data manipulation & numerical operations; pandas handle structured data (DataFrames), while NumPy supports array-based operations.
- b. **matplotlib, seaborn & plotly.express**: Libraries for data visualization; matplotlib and seaborn create static plots, while plotly enables interactive visualizations.
- c. **sklearn (scikit-learn)**: Supports end-to-end machine learning processes through tools for data cleaning, model training, and validation metrics like accuracy, confusion matrix, and ROC-AUC.
- d. **xgboost (XGBClassifier)**: A powerful and optimized gradient boosting library used for superior efficiency classification tasks.
- e. **imblearn (SMOTE)**: Handles imbalanced datasets using techniques like SMOTE to oversample minority classes.
- f. **scipy.stats**: Offers statistical tests (t-test, ANOVA, Chi-square) to analyze feature significance.
- g. **streamlit**: A framework to build interactive, shareable web apps tailored for showcasing data science and machine learning work.

## 9. Data Cleaning and Preprocessing:

We checked for gaps in data, as there are no missing values, we proceed with the further pre-processing techniques.

```
import pandas as pd

# Loading the dataset
df = pd.read_csv("heart.csv")
# Checking for missing values
print(df.isnull().sum())
```

```
Age          0
Sex          0
ChestPainType 0
RestingBP    0
Cholesterol  0
FastingBS    0
RestingECG   0
MaxHR        0
ExerciseAngina 0
Oldpeak      0
ST_Slope     0
HeartDisease 0
dtype: int64
```

Fig (8.1)

## One-Hot Coding:

```
import pandas as pd

# Loading the dataset
df = pd.read_csv("heart.csv")
df.columns = df.columns.str.strip().str.lower() #removing the whitespaces and converting the column names to lowercase

# Identifying the categorical columns from the dataset
categorical_cols = df.select_dtypes(include='object').columns.tolist()

# Applying one-hot encoding using pandas
df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=False)

# Displaying the output
print("Original shape:", df.shape)
print("Encoded shape:", df_encoded.shape)
df_encoded.head()

Original shape: (918, 12)
Encoded shape: (918, 21)
```

Fig (8.2)

To prepare our heart.csv dataset for ML models, we conducted **one-hot coding** to transform all ordinal features to continuous features. The available data includes several categorical attributes namely **Chest\_Pain\_Type**, **Sex**, **Resting\_ECG**, and **Incline of ST segment (Slope\_ST)**, which represent non-numeric values like categories or classes. As algorithms like

Logistic Regression, Random Forest, & XGBoost are designed to work with numerical input, categorical data has to be modified through proper encoding methods.

One-hot coding transforms each categorical feature into multiple binary (0 or 1) columns, where each new column represents a single unique category. For example, the feature ChestPainType, with values such as "typical angina" and "asymptomatic", would be split into distinct columns, each indicating the existence (1) or non-existence (0) of that category for a specific record.

This approach makes sure that:

- The model treats all categories as **equally weighted and unordered**, avoiding false assumptions about any implicit ranking among them.
- The input data becomes fully numeric, making it compatible with the mathematical operations performed by machine learning algorithms.
- It prevents the model from learning misleading relationships based on arbitrary numerical label assignments.

Thus, by applying one-hot encoding, we make the data suitable for training and ensure fair representation of all categorical values across the dataset.

### **Standard Scaler:**

Standard Scaler standardizes attributes by deducting the average & adjusting to a variance of 1, guaranteeing that every feature has a balanced impact on the models. This is especially important for algorithms influenced by the scale of input features, like logistic regression. After implementing this, we noticed the following changes:

- **Equal Weighting of Features:**  
After standardization, age (which ranges from ~30 to ~70) and cholesterol (~100–600) are on the same scale.  
The model now treats them equally during training, without favoring large-magnitude values.
- **Improved Convergence for Gradient-based Models:**  
Algorithms like Logistic Regression and XGBoost rely on gradient descent.  
Standardized data helps gradients allow for smoother training due to reduced unpredictability, enhancing both speed and accuracy.
- **Bias-Free Distance Calculation:**  
Normalization increases the predictive capability of NB and LR by making distance and probability estimates more consistent.
- **Better Handling of Multicollinearity:**
- The highly correlated variables can be reduced by giving them normal influence using standardization.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	\
0	-1.433140	M	ATA	0.410909	0.825070	0	Normal	
1	-0.478484	F	NAP	1.491752	-0.171961	0	Normal	
2	-1.751359	M	ATA	-0.129513	0.770188	0	ST	
3	-0.584556	F	ASY	0.302825	0.139040	0	Normal	
4	0.051881	M	NAP	0.951331	-0.034755	0	Normal	

	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	1.382928	N	-0.832432	Up	0
1	0.754157	N	0.105664	Flat	1
2	-1.525138	N	-0.832432	Up	0
3	-1.132156	Y	0.574711	Flat	1
4	-0.581981	N	-0.832432	Up	0

Fig (8.3)

We begin by loading the dataset and identifying quantitative variables - age\_group, BP, Cholesterol, MaxHR and Oldpeak. We apply StandardScaler from the sklearn.preprocessing module to enhance productivity efficiency, accuracy, & to ensure that the features are of comparable proportion.

StandardScaler standardizes numerical features by eliminating the average and normalizing to variance of 1. This change leads to every feature having an average of 0 and a statistical variability of 1. This procedure guarantees that attributes with higher numeric spans do not overshadow those with lower spans, which is particularly important for algorithms delicate to data standardization like logistic regression. By applying standard scaling, the model is better able to learn significant designs & connections within the data, resulting in enhanced proficiency and consistency throughout training.

## 10. Statistical Description:

```
Total patient records: 918
Patients with heart disease: 508
Patients with no heart disease: 410
```

Fig (10.1)

### Heart Disease Distribution:

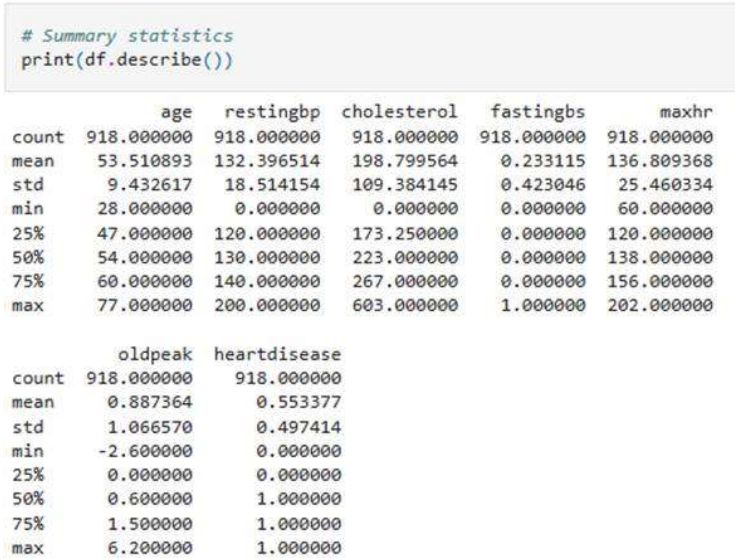
The data-set comprises two categories of patients, grounded in the existence or absence of HD. Patients diagnosed with heart disease are assigned a label of **1**, totaling **508 individuals**. Conversely, patients without HD are assigned a label of **0**, amounting to **410 individuals**.

### Slight Class Imbalance:

Out of a total of **918 patients**, approximately **55.3% (508 patients)** are diagnosed with heart disease (label = 1), while the remaining **44.7% (410 patients)** show no signs of heart disease (label = 0).

This indicates a **mild class imbalance**, which should be considered during model training, especially for classification algorithms (e.g., may affect precision/recall).

**Descriptive summary:**



**Fig (10.2)**

Column	Mean	Std Dev	Min	25%	50%	75%	Max
age	53.51	9.43	28	47	54	60	77
restingbp	132.40	18.51	0	120	130	140	200
cholesterol	198.80	109.38	0	173.25	223	267	603
fastingbs	0.23	0.42	0	0	0	0	1
maxhr	136.81	25.46	60	120	138	156	202
oldpeak	0.89	1.07	-2.6	0.0	0.6	1.5	6.2
heartdisease	0.55	0.50	0	0	1	1	1

**Interpretation by Feature:**

**Age:**

- Patients' age\_groups fall between **28 to 77**, with an average of around **53.5 years**.
- Most patients fall between **47 (25%) and 60 (75%)**, indicating a middle-aged population.

**Resting Blood Pressure (restingbp):**

- Mean is **132 mmHg**, which is borderline **hypertensive**.
- Min is **0**, which is **invalid** → suggests **data entry error**.
- Most values are between **120–140 mmHg**, expected for clinical patients.

**Cholesterol:**

- Mean is **198.8 mg/dL**, close to the borderline high level.
- **Minimum is 0**, which is also **invalid**, likely another data error.
- Max value **603** is extremely high.
- Consider **cleaning** this feature before modeling.

**Fasting Blood Sugar (fastingbs):**

- Binary variable: mostly **0** (no high sugar), with **23.3% of patients** having a fasting BS level greater than 120 mg/dL, which is represented by **1** in the dataset. It is **heavily skewed** toward 0.

**Peak Heart Rate(maxhr):**

- Falls around **60 to 202**, with an average of around **137 bpm**.
- Typically used in stress testing — may indicate cardiovascular capacity.

**Oldpeak:**

- This is likely **ST depression** (exercise-induced).
- Mean of **0.89** suggests mild depression.
- **Minimum value is -2.6**, which is **physiologically suspicious** — may need review.
- Max is **6.2**, indicating significant ischemia in some patients.

**Heart Disease Label:**

- Binary: 1 = disease, 0 = no disease.
- Mean of **0.553** means **~55.3% of patients have heart disease**.
- Median is 1 → **more than half of the data is +ve** for heart disease.

## Key Observations:

### 1. Data Quality Issues:

- restingbp and cholesterol have values of **0**, which are likely **incorrect**.
- oldpeak has **negative values**, which may be invalid depending on the context.

### 2. Possible Data Cleaning:

- Consider replacing or removing invalid zero/negative values.
- Check if these features have missing value codes that need to be treated.

### 3. Balanced Features:

- Binary columns like fastingbs and heartdisease show meaningful distributions.
- Most continuous features show reasonable spread for model training.

## 11. Exploratory Data Analysis:

### I. Acquisition of the Dataset:

This dataset was originally taken from across 4 different places, and we have visualized these places.

Location	Number of Records
Cleveland	303
Hungary	294
Switzerland	123
Long Beach VA	200
Stalog (Heart) Data Set	270

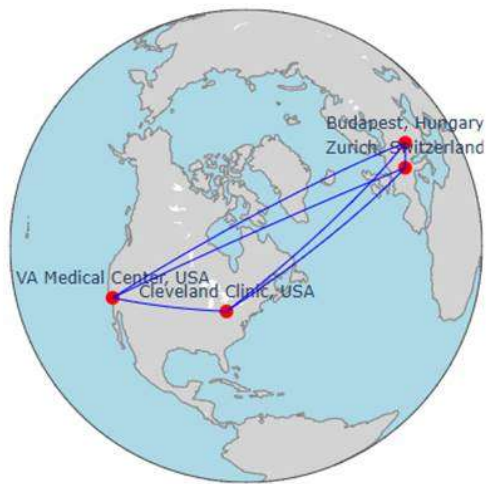


Fig (11.1)

## II. Univariate Analysis:

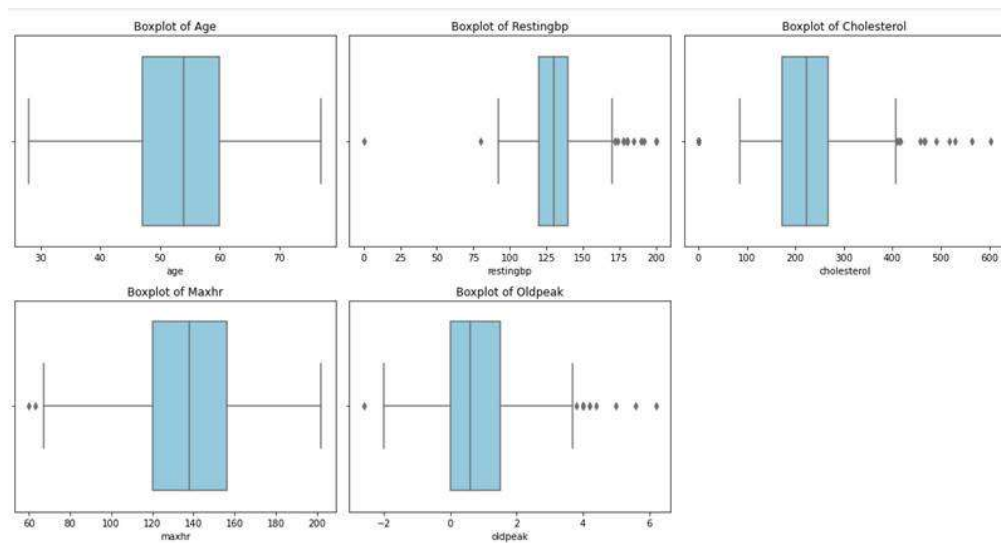


Fig (11.2)

### 1. Age

- **Distribution:** Symmetric, with no noticeable outliers.
- **Age range:** Mostly between ~40 to ~65 years.
- **Insight:** Age is well distributed, with minimal noise.



## 2. Resting Blood Pressure (restingbp)

- **Key Issue: Outliers on both ends**, especially near 0.
- **Problem:** A value of **0 mmHg is physiologically impossible**—likely a data error.
- **Other outliers:** Values above ~160 mmHg appear as right-side outliers.
- **Action:** Consider **imputing or removing zero values**; high values may reflect actual hypertensive patients.

## 3. Cholesterol

- **Outliers:** Extreme outliers on both ends — again, including **0 mg/dL**, which is likely invalid.
- **Long right tail:** Indicates **right-skewed distribution** with very high cholesterol values (>400–600).
- **Action:** Investigate **0 values** and consider **log-transforming** or capping extreme values.

## 4. Max Heart Rate (maxhr)

- **Outliers:** A few **low-end outliers** (e.g., <80 bpm), possibly representing abnormal conditions or measurement errors.
- **Distribution:** Slightly skewed left.
- **Insight:** Most people achieve ~120–160 bpm, which is expected.

## 5. Oldpeak (ST Depression)

- **Negative value:** At least one value is below 0 — which is **invalid** in most ST depression scales.
- **Outliers on the right:** Several patients with high ST depression (up to 6.2).
- **Insight:** High oldpeak values often indicate severe ischemia.

### III. Bivariate Analysis:

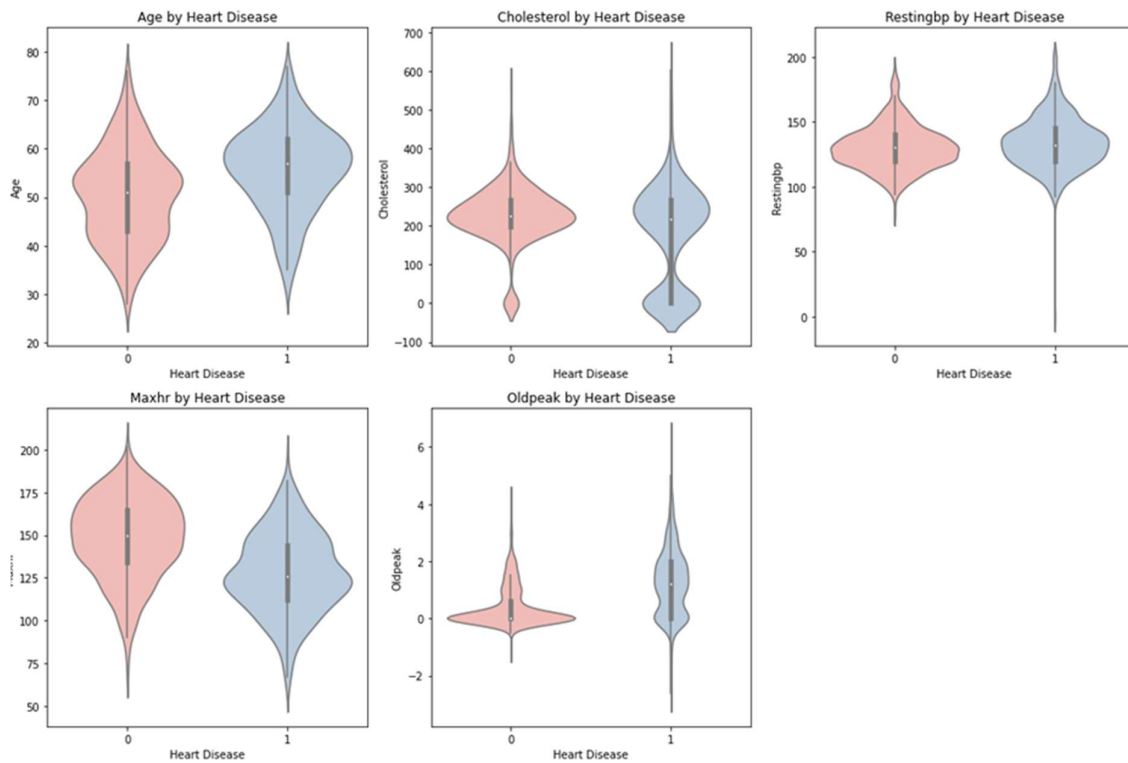


Fig (11.3)

#### Age-group

- Cardiac patients (1) are prone to be **older** on average.
- Age is concentrated around **50–65** in both groups.

#### Cholesterol

- Similar distributions for both classes, yet:
  - **Cardiac patients** show more **extreme low and high values**, including **invalid zeros**.
- Not strongly discriminatory on its own.

#### RestingBP

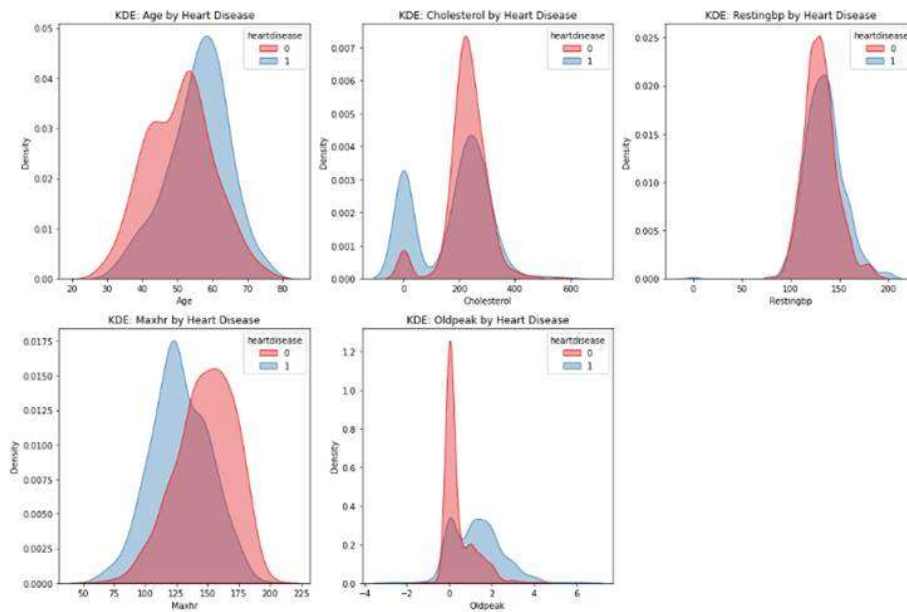
- Overlap exists, but the heart **disease group** slightly skews **higher in resting BP**.
- Some invalid low values (near zero) are visible.

#### MaxHR (Maximum heart\_rate)

- Non-cardiac patients generally reach **higher max heart-rates**.
- Lower maxHR in cardiac patients is consistent with reduced cardiac function.

### Oldpeak (ST Depression)

- **Strongest visual difference:**
  - Cardiac patients tend to have **higher oldpeak values**, indicating ischemia.
  - Patients without disease mostly have values near **0**.



**Fig (11.4)**

The KDE plot code is performing bivariate analysis using Kernel Density Estimation (KDE) to visualize the distribution of each numerical variable across the two classes of heart\_disease.

This KDE plot does:

1. It plots the smoothed probability distribution for individuals of cardiac and non-cardiac patients.
2. Different hues are applied on all class (0 = No, 1 = Yes) via `hue='heartdisease'`.
3. `fill=True` makes the curves easier to compare visually.
4. `common_norm=False` ensures each group is normalized independently for better comparison.

**Finding insights from the output: 1. Shape & Overlap** If the KDE curves for heart disease = 0 and 1 overlap heavily, the variable may not be a strong discriminator. Less overlap means the variable helps distinguish between the 2 classes.

**2. Peak Shifts** If one group has a peak at a higher or lower value, it indicates a trend: Example: oldpeak peaks at higher values for heart disease patients. Maxhr peaks at higher values for healthy individuals.

Feature	Distribution Insight	Predictive power
Age_group	Old patients tend to have Heart_Disease	Moderate
Cholesterol	Broad overlap; not a strong separator	Low
RestingBP	Overlapping curves; low separation	Low
MaxHR	Lower max HR seen in heart disease patients	Moderate to Strong
Oldpeak	Higher oldpeak values linked to heart_disease	Strong

#### IV. Bar Chart: Age vs Heart\_Disease:

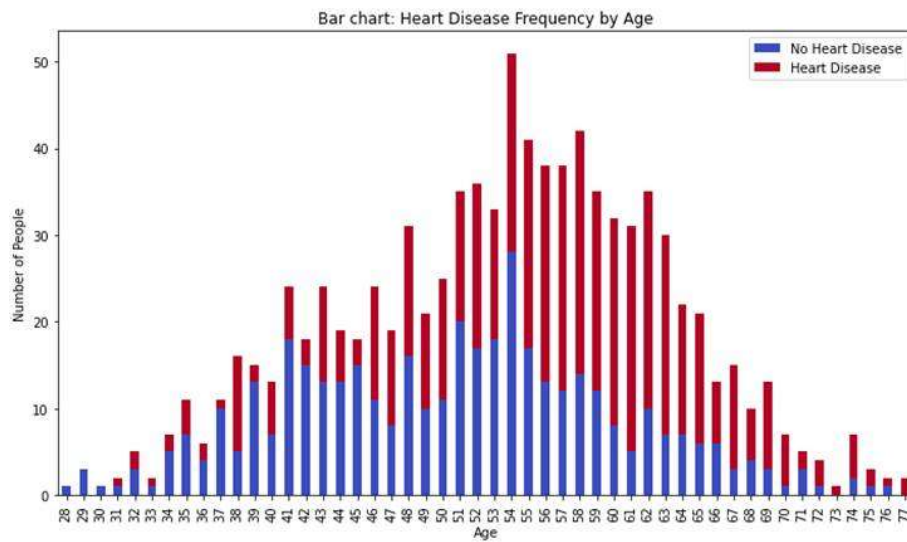


Figure (11.5)

##### Interpretation:

1. The X-axis shows age values (e.g., 28 old to 77 old).
2. The Y-axis shows the number of patients for all age\_group.

##### Bars are stacked with:

1. One color for cardiac patients (heartdisease = 0)
2. Another color for non-cardiac patients (heartdisease = 1)
3. The chart shows the incidence of heart disease segmented by age.

##### Insights drawn:

###### 1. Age Groups with Highest Cases:

Some age groups (e.g., 52, 54, 58, 60) may have taller bars, meaning more people in that age range are in the dataset. We can identify peak heart disease ages if the upper segment (disease = 1) dominates.

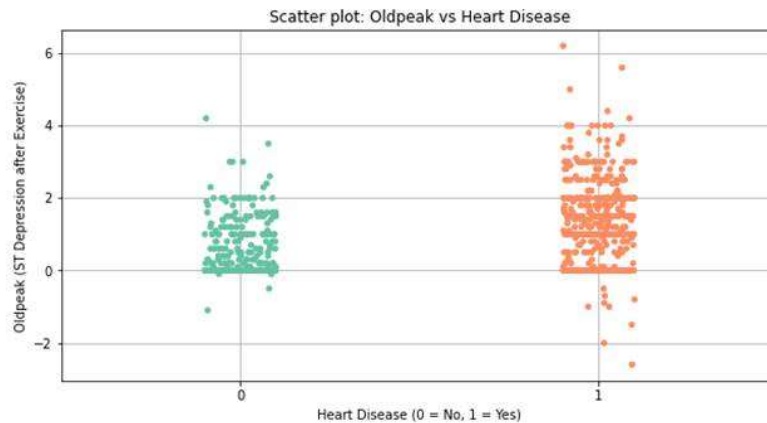
###### 2. Heart Disease Tends to Increase with Age:

In many datasets, you'll notice that older age groups (50–65+) have more red or dark-colored segments → indicating more cases of heart disease. Younger age groups (e.g., 30s, 40s) have smaller bars and are often more blue, indicating a lower possibility of heart disease.

### 3. Comparing Within Age Groups:

Within any specific age group, if the red segment (disease = 1) is taller than the blue segment, that age group has more patients with heart disease. Some mid-age groups might have balanced bars → indicating mixed risk.

### V. OldPeak vs Heart\_Disease:



**Fig (11.6)**

#### Interpretation:

##### Plot details:

1. x-axis: heart\_disease (binary: 0 = Non-cardiac, 1 = Cardiac)
2. y-axis: oldpeak values (ST depression after exercise)

Points: Each point is a patient, and their vertical position shows their oldpeak value.

jitter=True: Adds horizontal spread to prevent points from overlapping.

Color: Different hues (from Set2) for visual distinction.

#### Insights drawn:

##### 1. Patients Without Heart Disease (x = 0):

1. Most points are clustered near oldpeak = 0.
2. Very few high oldpeak values.
3. This suggests lower ST depression is prevalent in healthy patients.

##### 2. Patients With Heart Disease (x = 1):

1. Larger distribution of oldpeak values.
2. More patients have moderate to high oldpeak (e.g., values > 1.5).
3. Indicates greater ST depression shows a +ve relation to heartdisease.

VI. Violin plot: Age vs Chest pin type:

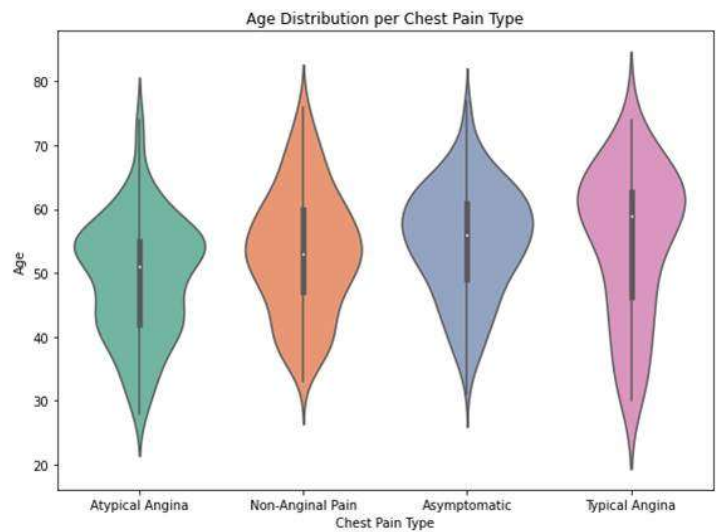


Fig (11.7)

Interpretation:

This violin plot visualizes how age is distributed across different types of chest pain, helping identify patterns in age-related chest pain presentations among heart patients.

Plot details:

**x-axis:** chestpaintype (Mapped categories: Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NTA), Asymptomatic (ASY))

**y-axis:** Patients age

**Violin Shape:** Shows distribution, density, and spread of age for each chest pain category.

Insights drawn:

ChestPain_Type	Age Distribution Insight
Typical Angina	More common in <b>older adults</b>
Atypical Angina	Occurs in <b>middle-aged</b> groups
Non-Anginal Pain	<b>A wide range</b> of ages was affected.
Asymptomatic	Seen mostly in <b>older adults</b> , it may indicate a silent risk.

VII. Beeswarm plot: Cholesterol vs heart disease:

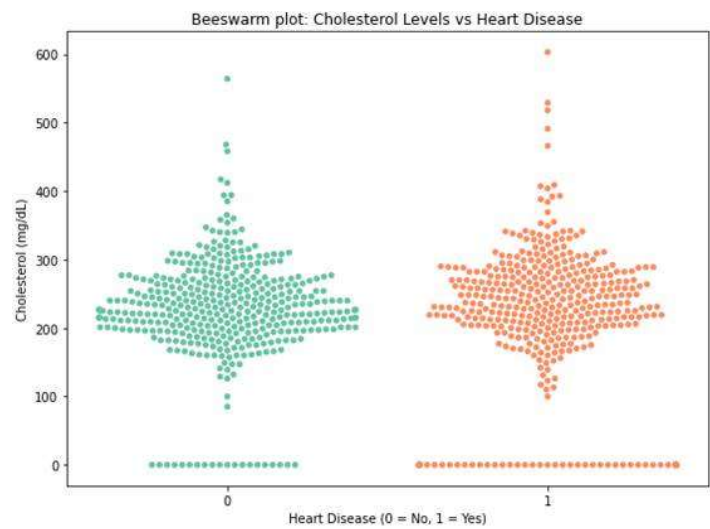


Fig (11.8)

Interpretation of output:

The bee swarm plot (a variation of a scatter plot with jittering) visualizes how lipid levels are distributed among individuals with & without HD.

Plot details:

x-axis: heart-disease

0 = No HD

1 = Has HD

y-axis: cholesterol (in mg/dL)

Each dot: A single patient's cholesterol level. Points are spread out horizontally to avoid overlap, giving a "bee swarm" look.

Insights drawn:

Observation	Interpretation
Wide overlap between classes 0 and 1	Cholesterol is <b>not a standalone discriminator</b>
High values in both classes	Some patients have high cholesterol but no disease.
Spread is more vertical than grouped.	Cholesterol varies widely within both groups.



## XII. Correlation Matrix:

### Interpretation:

This heatmap visualizes the correlation coefficients between all numeric features, helping identify relationships that are:

1. Significantly positive (values close to +1)
2. Significantly negative (values close to -1)
3. Weak or no correlation (values near 0)

The value shown at the intersection of two variables is the Pearson correlation coefficient.

### Color:

1. **Dark red** = significant +ve correlation
2. **Dark blue** = significant -ve correlation
3. **Lighter shades** = weak or no correlation

The diagonal is always **1.00** because each variable is perfectly correlated with itself.

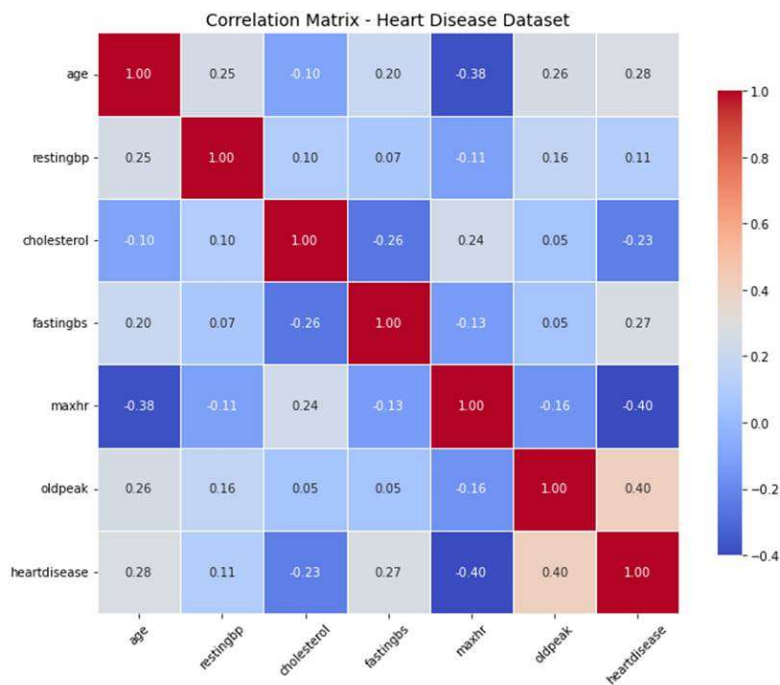


Fig (11.9)

### Observation:

The correlation analysis reveals key insights into different physiological characteristics and the occurrence of heart\_disease. Among all features, (maxHR) shows the strongest -ve correlation with HD (-0.40), suggesting that patients with low BPM are predicted to have heart disease.

This aligns with clinical expectations, as a diminished ability to reach higher heart rates during activity often reflects compromised cardiovascular performance. On the other hand, ST depression (oldpeak) depicts the strongest-positive-correlation (+0.40), which means the patients with greater ST depression are an indicator of myocardial ischemia and often have a higher likelihood of heart disease. Additional moderate correlations include age (+0.28) & **Pre-prandial glucose** (FastingBS) (+0.27), suggesting that elderly patients and those with increased fasting BS levels are at high danger. These patterns are consistent with established medical knowledge, where age and diabetes-related indicators are known risk factors. Meanwhile, cholesterol (−0.23) and resting blood pressure (+0.11) show weak or inconsistent correlations with heart disease, implying that on their own, they are not strong predictors in this dataset. These findings suggest that while traditional risk factors like cholesterol are still relevant, features like max heart rate and ST depression may offer more direct predictive value in identifying patients with heart disease in this particular dataset.

**Insights drawn:**

Feature	Correlation	Meaning
oldpeak	<b>+0.40</b>	High ST segment = heart disease
sex	<b>+0.30</b>	Males (usually coded as 1) are prone to having heart disease
chestpaintype	<b>+0.28</b>	Some types of chest pains are effectively linked with heart disease
FastingBS	<b>+0.26</b>	High fasting sugar levels may indicate heart risk
exerciseangina	<b>−0.49</b>	Presence of exercise-induced angina <b>reduces</b> heart disease likelihood (inverse)
maxhr	<b>−0.40</b>	Higher maximum heart rate = lower risk (healthy heart performance)
age	<b>−0.22</b>	Mild negative relation — older people are slightly less represented (data-specific)
cholesterol	<b>−0.06</b>	Very weak relationship — cholesterol alone isn't a good predictor here

## X. Sankey Diagram:

Chest Pain Type	Heart Disease	No Disease	Total
ASY	<b>392</b>	104	496
ATA	24	<b>149</b>	173
NAP	72	<b>131</b>	203
TA	20	26	46

Here is a **short explanation of each chest pain type** and its symptoms:

### 1. ASY (Asymptomatic)

- Meaning: No chest pain or noticeable symptoms.
- Symptoms: Often no warning signs; heart disease may go undetected until severe.
- Risk: Very high — commonly linked to silent heart attacks.

### 2. ATA (Atypical Angina)

- Meaning: Abnormal chest pressure not typical of heart-related pain.
- Symptoms: May include sharp, stabbing, or burning pain, often not triggered by exertion.
- Risk: Generally lower risk of heart disease.

### 3. NAP (non-Anginal Pain)

- Meaning: Chest\_pain is not associated with heart problems.
- Symptoms: Often due to muscle strain, acid reflux, or anxiety. Pain is localized and positional.
- Risk: Moderate — can mimic cardiac pain but usually isn't.

### 4. TA (Typical Angina)

- Meaning: Classic heart-related chest pain happens when the blood circulation to the heart is reduced.
- Symptoms: Compression, pressure, or heaviness in the chest, often triggered by physical exertion/depression.
- Risk: High — usually indicates underlying coronary artery disease.

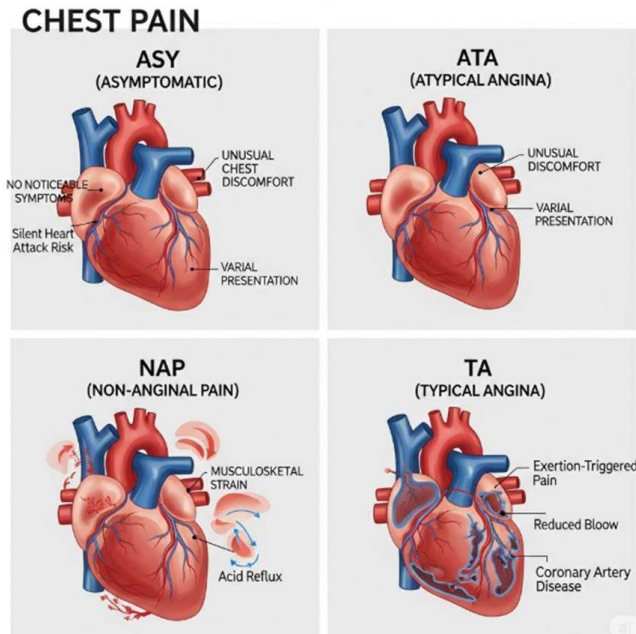


Fig (11.10)

#### Heart Disease Flows:

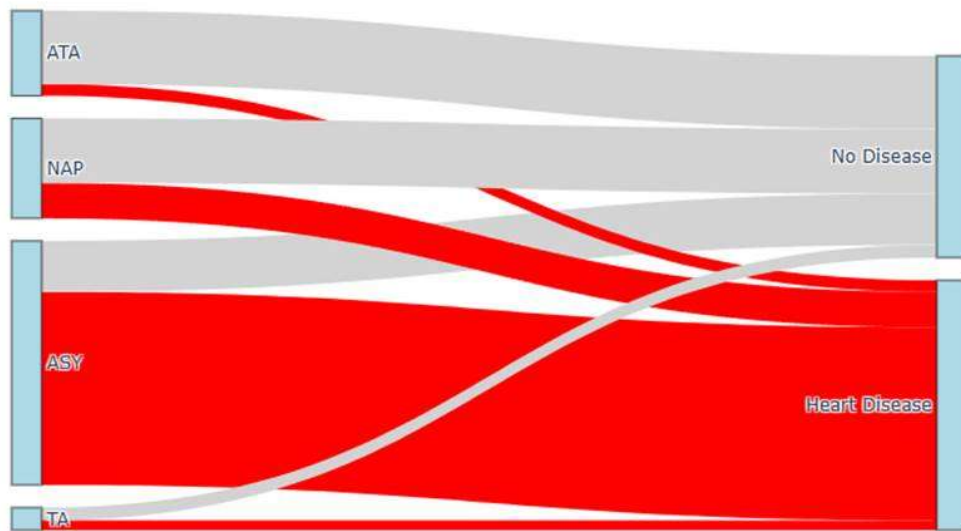
- **ASY (Asymptomatic)** contributes the most to heart disease: **392 patients**.
- **NAP (non-Anginal Pain)** contributes to 72 cases.
- **ATA (Atypical Angina)** contributes to 24 cases.
- **TA (Typical Angina)** contributes 20 cases.

This suggests that **asymptomatic individuals** (those not reporting chest\_pain) are prone to have HD—possibly because their condition goes unnoticed until it's severe.

#### No Disease Flows:

- **ATA (Atypical Angina)** is most common among patients **without heart disease: 149 individuals**.
- **NAP** follows with **131 no-disease cases**.
- **ASY** still has 104 cases with no disease.
- **TA** has 26 individuals without heart disease.

This shows that patients with **atypical or non-anginal pain** tend to be **free from Heart\_Disease**, meaning their symptoms are less predictive.



**Fig (11.11)**

The thickness of each flow (link) indicates the number of patients flowing from each chest pain category to the disease outcome.

**Heart Disease (Red Flows):**

- ASY (Asymptomatic) contributes the largest red flow, indicating that the most of the patients with asymptomatic chest discomfort are detected with HD. This aligns with the risk of silent or undetected conditions, making it a critical warning signal.
- NAP (Non-Anginal Pain) and TA (Typical Angina) also show moderate flows toward heart disease, indicating a notable portion of patients with these symptoms are affected.
- ATA (Atypical Angina) has the smallest red flow, suggesting it's less likely to be linked with HD.

**No Disease (Gray Flows):**

- ATA (Atypical Angina) sends a thick gray flow to "No Disease", meaning many people with this symptom do not have heart disease.
- NAP also shows a large gray flow, indicating that this symptom often occurs without heart disease.
- ASY still has some flow to the "No Disease" side, but much smaller than its red flow—confirming it's highly predictive of disease.

### Observation:

1. ASY is a strong predictor of heart disease — even though it’s “asymptomatic,” these individuals silently carry high risk.
2. ATA and NAP are more prevalent in patients without heart disease, making them less predictive symptoms.
3. The diagram effectively shows how symptom type influences diagnosis, helping clinicians target high-risk patients.

## XII. Simulated ECG Plot based on RestingECG, Oldpeak, maxHR:

This plot displays ECG signals for four random patients, using data from a heart disease dataset. It generates and visualizes these ECG-like waveforms by factoring in:

- **maxhr** (high heart rate),
- **oldpeak** (exercise provoked-ST depression), &
- **restingecg** (ECG result at rest: Normal, LVH, or ST).

```
generate_ecg_signal(heart_rate, oldpeak, rest_ecg, length=1.0, fs=250)
```

Fig (11.12)

The waveform is shaped:

- **t (time axis):** 0 to 1 second, sampled at 250 Hz.
- **baseline:** A decaying sinusoid that mimics heartbeats, shaped by:
  - Heart rate (heartrate) → determines frequency.
  - oldpeak → depresses the waveform (simulating ST depression).
  - rest\_ecg:
    - 'LVH' → adds additional frequency to simulate strain.
    - 'ST' → shifts voltage upward (simulating ST elevation).

Each plot represents an **ECG shape**, not a real clinical ECG. Here is how to read them:

- **Higher heart rate** → more frequent oscillations.
- **Presence of oldpeak** → waveform dips downward more (depression).
- **restingecg = 'LVH'** → waveform becomes more complex (added high-frequency component).

- **Disease:**
  - You may observe flatter or irregular shapes in red plots.
  - Green plots tend to look smoother and more regular.

### Example Interpretation of a Single Plot:

#### Title: Patient 535 | Disease

- Red waveform.
- Shows an irregular pattern, likely flattened due to oldpeak.
- Possibly has high-frequency spikes if `restingecg == LVH`.

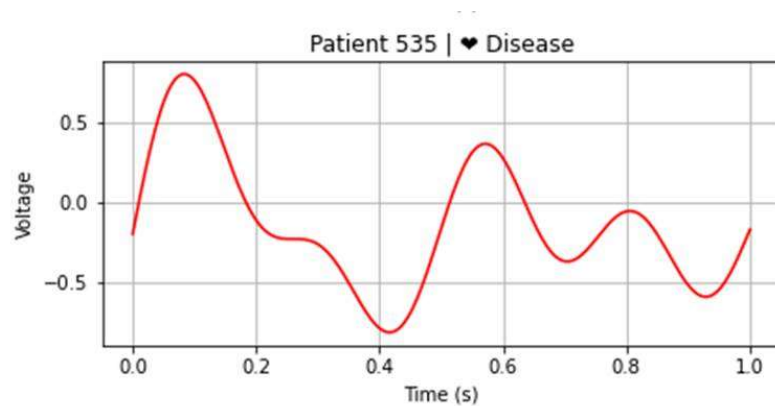


Fig (11.13)

#### Title: Patient 668 | No Disease

- Green waveform.
- Smooth, regular sinusoidal pattern.
- Indicates a relatively healthy simulated ECG pattern.

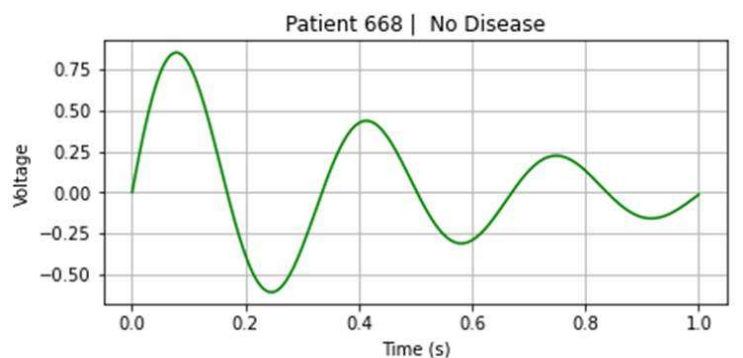


Fig (11.14)

This visual representation shows:

- Visually contrasts ECG-like signals between patients with and without heart disease.
- Incorporates real patient data to simulate changes based on heart rate and ECG findings.
- While not medically accurate, it gives a visually intuitive and informative representation for educational and demonstration purposes.

## XII. Importance Severity Hazard (ISH) Chart:

The ISH chart is a **visual diagnostic tool** to analyze medical features (like BP, lipid level, chest pain\_type, etc.) in terms of their relationship with heart disease.

Term	Meaning
<b>Importance</b>	How strongly a feature is correlated with heart disease (Pearson's correlation).
<b>Severity</b>	How much the feature varies in the dataset (Standard Deviation).
<b>Hazard</b>	How often does the feature show <b>extreme values or outliers</b> , based on IQR.

### Plot Overview:

- **X-axis (Importance):**
  - a) Features to the **right** are more positively correlated with heart disease.
  - b) Features to the **left** are more negatively correlated.
  - c) Features near 0 have weak or no correlation.
- **Y-axis (Severity):**
  - a) Features **higher up** vary more among the population (i.e., standard deviation is high).
  - b) Low severity means consistent values across patients.
- **Bubble size (and color): Hazard**
  - a) Larger and **darker red** bubbles indicate more extreme values (outliers) in the population.
  - b) Calculated using the IQR method (features with lots of outliers have higher hazard).
- **Text labels:** Feature names placed near each bubble for easy identification.



### Interpretation of the output:

This visualization allows us to **triage features** based on their **medical risk profile**:

#### ● High Importance + High Hazard + High Severity

- **Most critical features.**
- Strongly associated with heart disease, vary a lot, and have many outliers.
- **Example:** If Cholesterol or Oldpeak is in this zone, → clinically significant.

#### ● High Importance + Low Hazard

- Consistently important indicators, even if values are not extreme.
- Still valuable for prediction.

#### ● Low Importance + High Hazard

- May not be useful for prediction, but could represent risk-prone behavior or measurement errors.

#### ○ Near Origin (0,0)

- Not useful as risk indicators—neither important nor variable.

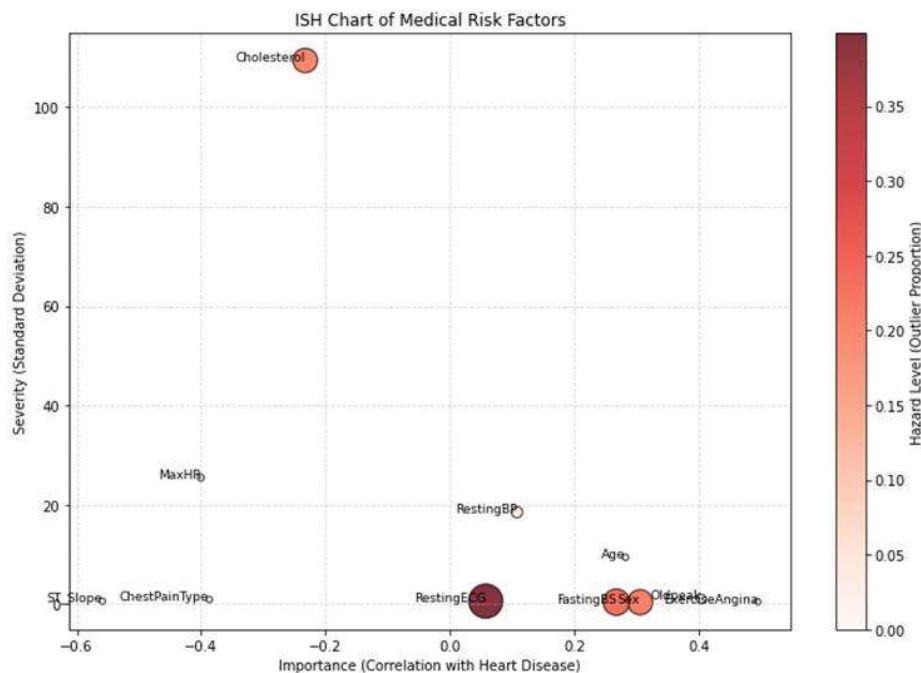


Fig (11.15)

From the chart:

- Exercise Angina, Oldpeak, and FastingBS stand out with high importance, meaning they are highly linked with the confirmed case of heart disease. Among them, Oldpeak and FastingBS also show moderate to high hazard levels (darker shades), indicating that extreme values in these features are more common and possibly clinically significant.
- Cholesterol shows a moderately negative correlation with heart disease (importance  $\approx -0.2$ ), but it has very high severity, meaning that cholesterol values vary widely across patients. Its hazard level is also notable, suggesting that some patients exhibit extreme cholesterol levels, which may still be clinically relevant even if correlation is low.
- RestingECG is near the center but shows a high hazard level (large, dark bubble), indicating many outliers despite having a weak correlation. This could suggest inconsistencies or subgroup-specific significance in the ECG results.
- Features like ChestPainType and ST\_Slope appear on the far left with negative correlation and low variability and low hazard, suggesting they may be less useful in predictive modeling or less clinically significant for this dataset.
- MaxHR and RestingBP show moderate severity but low correlation and low hazard, indicating they vary but aren't strongly predictive of disease here.

In summary, features to prioritize in modeling or clinical screening include Exercise Angina, Oldpeak, and FastingBS for their high correlation and hazard, while features like Cholesterol warrant attention due to their high variability and moderate hazard, even if their correlation is not strongly positive.

## 12. Feature Engineering for Exploratory Analysis:

In this script, we enhanced a heart disease dataset by engineering new features to better capture health risk patterns. We first cleaned the column names and converted binary categorical variables (fastingbs and exerciseangina) to numeric format. Then, we categorized patients into **age groups** (Young, Middle-aged, Senior). A **composite metabolic risk** score was created by normalizing cholesterol and combining it with fasting blood sugar. Finally, we flagged patients under **severe stress** if their oldpeak was above 2 and they experienced exercise-induced angina. These new features help in improving both clinical interpretation and predictive modeling.

### Steps taken:

1. We converted the categorical features into binary features for e.g: converting text labels ('Y', 'N') into numeric binary (1, 0) values.
2. We created a new categorical feature called age\_group, which makes it easier to analyze heart disease prevalence by life stage.
3. We normalized cholesterol between 0 and 1. We combine it with fasting to create a composite risk factor called metabolic\_risk. It helps to flag patients with high cholesterol and high BP.

- We flag patients experiencing exercise-induced stress AND ST depression > 2. These conditions together may indicate critical heart risk. `severe_stress = 1` means the patient likely needs further attention.

We then finally print the new columns:

- `age_group`
- `cholesterol_norm`
- `metabolic_risk`
- `severe_stress`

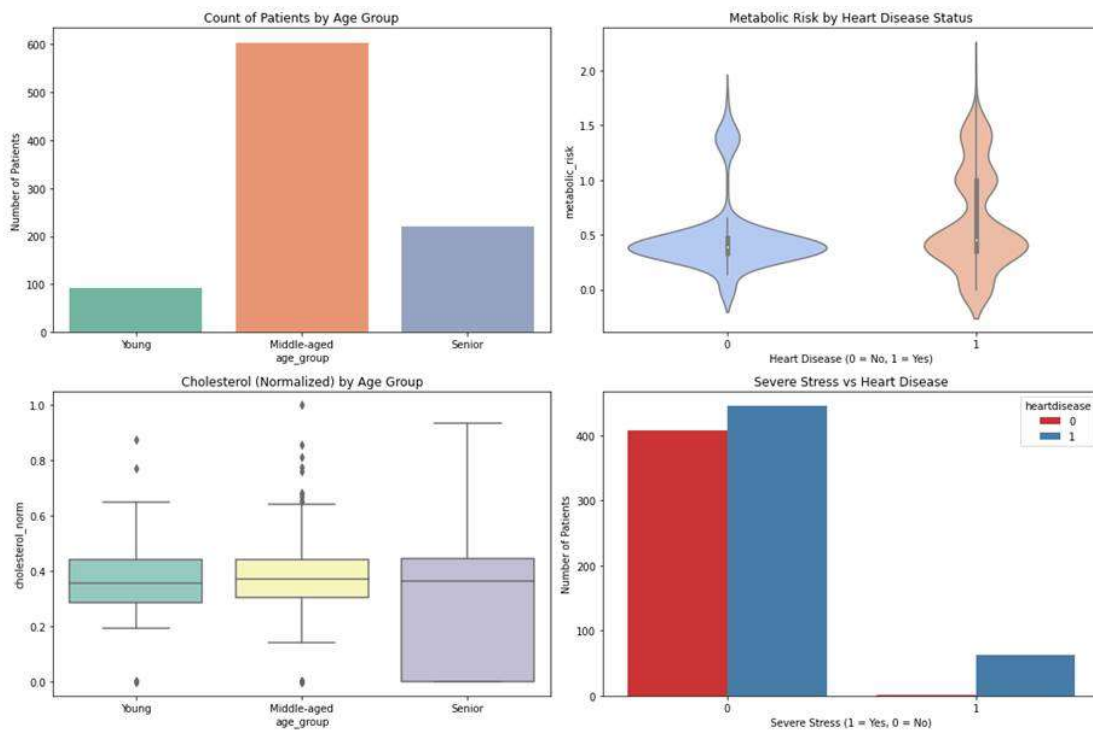
```

age    age_group  cholesterol  fastingbs  metabolic_risk  oldpeak
0     40    Young         289          0         0.479270      0.0
1     49  Middle-aged     180          0         0.298507      1.0
2     37    Young         283          0         0.469320      0.0
3     48  Middle-aged     214          0         0.354892      1.5
4     54  Middle-aged     195          0         0.323383      0.0

exerciseangina  severe_stress
0               0              0
1               0              0
2               0              0
3               1              0
4               0              0

```

**Fig (12.1)**



**Fig (12.2)**

## IMPORTANT NOTE:

“These features were engineered during EDA to explore patterns in the data. However, they were not included in model training.”

### 13. Feature Importance Analysis:

Here, we applied a Random Forest algorithm to perform heart disease prediction and used it to analyze feature importance. Categorical features were label-encoded to numeric form, & the data-set was split into train & test classes. Later after model training, we extracted and ranked the importance of each feature based on how much it contributed to the prediction, helping us find out the paramount factors allied to heart disease.

ST_Slope	0.242766
Cholesterol	0.113802
Oldpeak	0.108979
ExerciseAngina	0.103862
MaxHR	0.100141
ChestPainType	0.091722
Age	0.089734
RestingBP	0.069247
Sex	0.035929
RestingECG	0.025776
FastingBS	0.018043
dtype: float64	

Fig (13.1)

#### Interpretation:

- **Top Contributors:**

- a) ST\_Slope (0.242): Most important feature — the gradient of the peak exercise ST segment is highly predictive of HD.
- b) Oldpeak (0.108) and MaxHR (0.100): Indicators of exercise-induced abnormalities and heart rate response are also key predictors.
- c) Cholesterol (0.113) and ChestPainType (0.091): These are strong clinical indicators as well, aligned with traditional risk assessment.

- **Moderate Importance:**

- a) Age (0.089), ExerciseAngina (0.103), and RestingBP (0.025) — all relevant but slightly less impactful.

- **Low Importance:**

- a) Sex (0.035), RestingECG (0.025), and FastingBS (0.018) show lower predictive value in this dataset, suggesting they contribute less to the model's decisions.

#### Summary:

The model suggests that exercise-related variables (ST\_Slope, Oldpeak, Exercise\_angina, and MaxHR) are the key predictors of HD in this data-set, while basic demographics and fasting blood sugar have relatively low influence.

## 14. Hypothesis Testing:

### 1. Two-Sample Hypothesis Test:

We have 5 quantitative variables in our dataset, they are:

Column Name	Description
age	Age in years
restingbp	Resting blood pressure (in mm Hg)
cholesterol	Serum cholesterol (in mg/dL)
maxhr	Maximum heart rate achieved
oldpeak	ST depression induced by exercise

**Fig (14.1)**

In this analysis, we have performed independent two-sample t-tests to investigate whether five quantitative (continuous) features differ significantly between 2 classes: cardiac patients (HeartDisease = 1) and non-cardiac patients (HeartDisease = 0).

### Interpretation:

#### 1. Group Separation:

The dataset is divided into two groups according to the HeartDisease column:

- **group\_0**: Non-cardiac
- **group\_1**: Cardiac

This separation is essential for performing group-wise comparisons.

#### 2. Selection of Quantitative Features:

We selected five continuous variables to test:

- Age of patients
- RestingBP
- Cholesterol
- MaxHR
- Oldpeak

These are medically relevant features likely to show differences between diseased and non-diseased groups.

### 3. Performing the t-tests:

For each feature, the independent t-test (ttest\_ind) is applied. This test evaluates whether the means of two independent groups (with vs. without heart disease) are statistically significantly different.

- The **Null Hypothesis (H<sub>0</sub>)** for each test is:  
*There is no difference in the mean of the feature across the groups*
- The **Alternative Hypothesis (H<sub>1</sub>)** is:  
*There is a difference in the mean of the feature across the groups*
- nan\_policy='omit' ensures that missing values (if any) don't disrupt the calculation.

### 4. Output:

For every feature, the code produces the results of:

- **t-statistic:** Evaluates the difference size relative to the standard deviation.
- **p-value:** Tells you if the result is statistically significant. A (p<0.05) indicates a significant distinction in the feature between groups.

#### Output:

T-Test Results (Numerical Features vs Heart Disease)		
Feature	t-statistic	p-value
-----		
Age	-8.9	0.0
RestingBP	-3.28	0.001
Cholesterol	7.24	0.0
MaxHR	13.23	0.0
Oldpeak	-13.36	0.0

Fig (14.2)

Feature	t-statistic	p-value	Significant? (p < 0.05)	Interpretation
Age	-8.90	0.000	✔ Yes	Heart disease patients are <b>significantly older</b> than non-heart patients.
RestingBP	-3.28	0.001	✔ Yes	Resting blood pressure is <b>significantly lower</b> in heart disease patients.
Cholesterol	7.24	0.000	✔ Yes	Heart disease patients have <b>significantly higher cholesterol</b> levels.
MaxHR	13.23	0.000	✔ Yes	Maximum heart rate is <b>significantly higher</b> in heart disease patients. (Unusual; may need to check data)
Oldpeak	-13.36	0.000	✔ Yes	Oldpeak is <b>significantly lower</b> in heart disease patients. (Also unusual, usually it increases)

Fig (14.3)

#### Observation:

All features show statistically significant differences (p < 0.05), indicating meaningful group differences. Cardiac patients are older and have lower resting blood pressure, and high lipid levels on an average. Interestingly, they also have a higher max. HR and lower oldpeak

values, which is counterintuitive and may suggest either a data anomaly or unique population characteristics. These findings highlight which clinical variables most differentiate the two patient groups.

## 2. Chi<sup>2</sup> Test:

We have 5 categorical variables in our dataset, they are:

Column Name	Description
sex	Biological sex of the patient ( Male , Female )
chestpaintype	Type of chest pain: <ul style="list-style-type: none"><li>• TA – Typical Angina</li><li>• ATA – Atypical Angina</li><li>• NAP – Non-Anginal Pain</li><li>• ASY – Asymptomatic</li></ul>
restingecg	Resting electrocardiogram results: <ul style="list-style-type: none"><li>• Normal – Normal ECG</li><li>• ST – ST-T wave abnormality</li><li>• LVH – Left ventricular hypertrophy</li></ul>
exerciseangina	Exercise-induced angina: <ul style="list-style-type: none"><li>• 1 – Yes</li><li>• 0 – No</li></ul>
st_slope	Slope of the peak exercise ST segment: <ul style="list-style-type: none"><li>• Up – Upsloping</li><li>• Flat – Flat</li><li>• Down – Downsloping</li></ul>

**Fig (14.4)**

In this analysis, the Chi<sup>2</sup> Test was carried out to evaluate whether there is a statistically meaningful connection between the five categorical variables and the presence of heart\_disease.

### Insight:

#### Examined Categorical Variables:

- sex
- chest\_pain\_type
- RestingECG
- exercise\_angina
- ST\_slope

#### For every feature:

- A contingency table was created to compare the distribution of that variable across heart disease outcomes.
- Then applied the Chi-Square test using chi2\_contingency to evaluate independence.

#### Performing Chi<sup>2</sup> test:

For each feature, a Chi<sup>2</sup> test is applied to evaluate whether there exists a notable connection between each categorical feature and the heart disease outcome.

- The **null hypothesis ( $H_0$ )** for each test is:  
*The feature is independent of heart disease (no association).*
- The **Alternate Hypothesis ( $H_1$ )** is:  
*The feature is associated with heart disease.*

#### Output:

- The Chi-Square statistic tells us how far the observed frequencies deviate from expected frequencies under the null hypothesis.
- The p tells us whether this difference is statistically significant.
- A ( $p < 0.05$ ) indicates a significant association between the variable and heart disease, so we reject the null hypothesis (which assumes no association).

```
Chi-Square Test Results:

Variable: sex
Chi2 Statistic = 84.145, p-value = 0.0000, Degrees of Freedom = 1
➡ Significant association with Heart Disease (Reject H0)

Variable: chestpaintype
Chi2 Statistic = 268.067, p-value = 0.0000, Degrees of Freedom = 3
➡ Significant association with Heart Disease (Reject H0)

Variable: restingecg
Chi2 Statistic = 10.931, p-value = 0.0042, Degrees of Freedom = 2
➡ Significant association with Heart Disease (Reject H0)

Variable: exerciseangina
Chi2 Statistic = 222.259, p-value = 0.0000, Degrees of Freedom = 1
➡ Significant association with Heart Disease (Reject H0)

Variable: st_slope
Chi2 Statistic = 355.918, p-value = 0.0000, Degrees of Freedom = 2
➡ Significant association with Heart Disease (Reject H0)
```

Fig (14.5)

Variable	Chi <sup>2</sup> Statistic	p-value	df	Significant? (p < 0.05)	Interpretation
Sex	84.15	0.0000	1	✔ Yes	There is a <b>significant association</b> between sex and heart disease.
ChestPainType	268.07	0.0000	3	✔ Yes	Chest pain type is <b>strongly associated</b> with heart disease.
RestingECG	10.93	0.0042	2	✔ Yes	Resting ECG results are <b>significantly related</b> to heart disease.
ExerciseAngina	222.26	0.0000	1	✔ Yes	Exercise-induced angina is <b>highly associated</b> with heart disease.
ST_Slope	355.92	0.0000	2	✔ Yes	ST segment slope shows a <b>very strong association</b> with heart disease.

Fig (14.6)

#### Observation:

The Chi-Square test results show that all 5 ordinal variables—Sex, ChestPainType, RestingECG, ExerciseAngina, and ST\_Slope—show statistically significant associations with heart disease ( $p < 0.05$ ). Among them, ST\_Slope and ChestPainType exhibit the strongest



associations, indicating they may play a critical role in diagnosing HD. ExerciseAngina and Sex also show strong links, while RestingECG, though still significant, has a relatively weaker association. These findings suggest that categorical features meaningfully differ the patients between presence and absence of heart\_disease.

## 15. Contingency Table for all categorical variable's vs HD:

Contingency Table: sex vs heartdisease				
	sex	heartdisease = 0	heartdisease = 1	total
0	F	143	50	193
1	M	267	458	725
2	All	410	508	918

Contingency Table: chestpaintype vs heartdisease				
	chestpaintype	heartdisease = 0	heartdisease = 1	total
0	ASY	104	392	496
1	ATA	149	24	173
2	NAP	131	72	203
3	TA	26	20	46
4	All	410	508	918

Contingency Table: restingecg vs heartdisease				
	restingecg	heartdisease = 0	heartdisease = 1	total
0	Lvh	82	106	188
1	Normal	267	285	552
2	ST	61	117	178
3	All	410	508	918

Contingency Table: exerciseangina vs heartdisease				
	exerciseangina	heartdisease = 0	heartdisease = 1	total
0	N	355	192	547
1	Y	55	316	371
2	All	410	508	918

Contingency Table: st_slope vs heartdisease				
	st_slope	heartdisease = 0	heartdisease = 1	total
0	Down	14	49	63
1	Flat	79	381	460
2	Up	317	78	395
3	All	410	508	918

Fig (15.1)

## Observations:

### 1. Sex vs Heart-Disease:

Out of 918 patients, male (725) & female (193). A significantly higher number of males (458/ 508) have heart disease compared to females (only 50/193), suggesting that male patients are likely to have heart-disease in this dataset.

### 2. ChestPainType vs Heart-Disease:

The 'ASY' (asymptomatic) group is the most common (392 out of 508), while 'ATA' and 'NAP' types are more common among non-heart disease patients. This indicates a strong link between asymptomatic chest-pain & HD

### 3. Resting ECG vs Heart-Disease:

Patients with an ST ECG result show a higher number of HD cases (117 out of 178), while 'Normal' ECG is more balanced (285 vs 267). This suggests abnormal ECG readings are associated with heart disease.

### 4. Exercise Angina vs Heart-Disease:

Among those with exercise-induced angina ('Y'), 316 out of 371 have heart disease. On the other hand, most people without angina ('N') don't have the disease. This shows a strong relation between exercise-induced angina and heart-disease.

### 5. ST\_slope vs Heart-Disease:

Most people with heart\_disease show a 'Flat' ST segment (381 out of 460 cases), while the 'Up' slope is more common in those who do not have the condition. This indicates a significant correlation between a flat ST slope & the presence of HD.

## Insights:

- The existence of asymptomatic chest-pain (ASY), a flat ST segment, and exercise-induced angina are the influential categorical indicators of HD.
- In specific, patients showing a flat ST slope (81.1% have heart disease), reporting exercise-induced angina (85.2% show presence of HD), or presenting with asymptomatic chest pain (79% have HD) are significantly prone to have HD.
- This suggests that even in the absence of typical chest pain symptoms, subtle indicators like ECG patterns and exercise response play a crucial role in diagnosing heart disease early.

## 16. Anomaly Detection:

Detected 192 numerical anomalies based on clinical/statistical thresholds.

age	sex	chestpain	type	restingbp	cholesterol	fastingbs	restingecg	maxhr	exerciseangina	oldpeak	st_slope	heartdisease
28	53	F	ATA	113	468	0	Normal	127	0	0.0	Up	0
30	53	M	NAP	145	518	0	Normal	130	0	0.0	Flat	1
69	44	M	ASY	150	412	0	Normal	170	0	0.0	Up	0
76	32	M	ASY	118	529	0	Normal	130	0	0.0	Flat	1
98	56	M	ASY	120	85	0	Normal	140	0	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
624	63	F	ASY	150	407	0	LVH	154	0	4.0	Flat	1
667	65	F	NAP	140	417	1	LVH	157	0	0.8	Up	0
796	56	F	ASY	134	409	0	LVH	150	1	1.9	Flat	1
829	29	M	ATA	130	204	0	LVH	202	0	0.0	Up	0
850	62	F	ASY	160	164	0	LVH	145	0	6.2	Down	1

192 rows × 12 columns

Fig (16.1)

192 numerical anomalies were detected in the dataset based on clinical or statistical thresholds. These are patient records where one or more feature values deviate significantly from what is typically expected. The anomalies may reflect outliers or data quality issues, or they could represent rare but real clinical cases. After handling the anomalies using the IQR method, these are the results we got.

Original rows: 918  
After outlier removal: 701

Fig (16.2)

In this project, the IQR (Interquartile Range) method is ideal for handling outliers because it is robust and does not assume a normal distribution—making it well-suited for clinical data like cholesterol and oldpeak, which can be skewed. It helps detect truly unusual or potentially erroneous values without being affected by extreme cases. This method ensures that rare but important medical conditions are not wrongly removed, preserving the integrity of the dataset while improving model performance and reliability. Here are the **medical causes** of outliers observed in each numerical attribute in the Heart\_Disease dataset, based on clinical and physiological understanding:

### 1. Age

▲ **Outliers:** Typically, patients aged above 75 or below 30.

#### Medical Causes:

- **Younger Patients (<30):** Rare genetic heart conditions (e.g., congenital heart defects, familial hypercholesterolemia).
- **Older Patients (>75):** Increased cardiovascular risk due to aging, atherosclerosis, and cumulative health deterioration.

## 2. RestingBP (Resting Blood Pressure)

▲ **Outliers:** Blood pressure values >180 mm Hg or very low (<90 mm Hg).

### Medical Causes:

- **High BP:** Hypertension Stage 2 or hypertensive crisis, often linked to kidney disease, obesity, diabetes, or medication non-compliance.
- **Low BP:** Orthostatic hypotension, adrenal insufficiency, or heart valve problems.

## 3. MaxHR (Maximum Heart Rate Achieved)

▲ **Outliers:** HR <90 bpm or >190 bpm during stress testing.

### Medical Causes:

- **Low MaxHR:** Chronotropic incompetence (heart's inability to increase rate), heart block, medications like beta-blockers.
- **High MaxHR:** Arrhythmias, overactive thyroid (hyperthyroidism), or high physical fitness in young individuals.

## 4. Oldpeak (ST Depression Induced by Exercise)

▲ **Outliers:** Values >3.0 indicate severe ST depression.

### Medical Causes:

- **High Oldpeak:** Indicates significant myocardial ischemia or coronary artery disease (CAD), poor oxygen supply to the heart muscle.
- **Outlier values** may also reflect poor ECG calibration or stress testing errors.

## 5. Cholesterol

▲ **Outliers:** Total cholesterol >400 mg/dL or <100 mg/dL.

### Medical Causes:

- **Very High Cholesterol:** Familial hypercholesterolemia, poorly controlled diabetes, hypothyroidism, or high-fat diets.
- **Very Low Cholesterol:** Malnutrition, liver disease, or hyperthyroidism — all conditions where lipid synthesis is impaired.

These outliers often reflect serious underlying medical conditions, and while they may be removed for machine learning consistency, they can also offer clinically valuable information when analyzed separately. Outliers in this heart dataset can be due to medical conditions (e.g., high cholesterol, abnormal heart rate) or data entry errors (e.g., 0 values, negative numbers). Values within clinical range are likely genuine, while biologically implausible ones suggest recording mistakes.

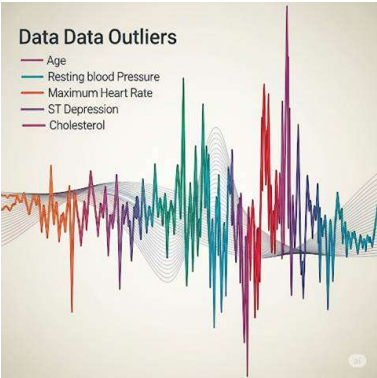


Fig (16.3)

**Visual Representation of numerical outliers and handling outliers:**

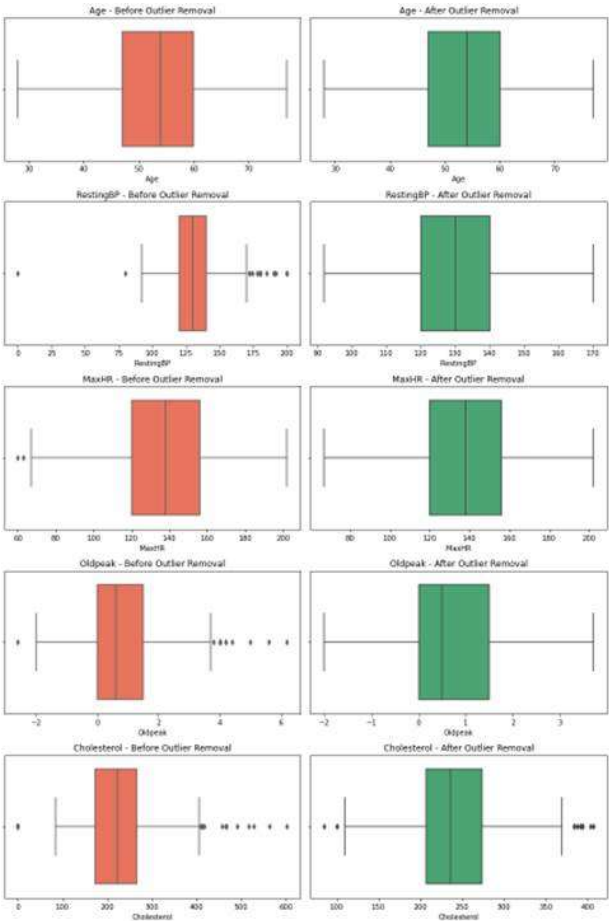


Fig (16.4)

## Observations:

### 1. Age

Before outlier removal, the age distribution appears reasonably compact with no extreme values. After removal, the boxplot remains largely unchanged, indicating minimal outliers in age data. This suggests age values were already clean and mostly within the expected clinical range.

### 2. RestingBP (Resting Blood Pressure)

The boxplot before shows multiple outliers on the higher end (above 180 mm Hg), indicating a few patients with unusually high blood pressure. After applying IQR, these outliers are removed, resulting in a cleaner, more symmetric distribution, helping improve model robustness.

### 3. MaxHR (Maximum Heart Rate)

Several outliers were present on both lower and higher ends, especially below 100 and above 180 bpm. After outlier removal, the distribution becomes more centered and consistent with physiological expectations, improving data reliability and reducing noise.

### 4. Oldpeak (ST Depression)

The original plot shows significant outliers above 2.5, which are clinically rare and may skew model predictions. Post-removal, the spread becomes tighter and more focused around common values (0–1), enhancing the accuracy of statistical analyses.

### 5. Cholesterol

This feature had the most extreme outliers, with values exceeding 500 mg/dL. These can mislead the model, as such high levels are rare or possibly data entry errors. After IQR removal, the distribution becomes much more normalized and medically plausible.

These changes demonstrate that the IQR method effectively reduces noise, making the data more consistent with real-world clinical thresholds and ultimately supporting better model training and interpretation.

## Anomaly Detection of Categorical Features:

```
No rare categorical patterns detected.  
NOTE:  
No rare/outliers have been detected in the categorical features.
```

**Fig (16.5)**

This code identifies rare categorical values (those occurring in less than 5% of the data) within selected columns of a heart disease dataset. It first standardizes column names, encodes binary categorical features like `exercise_angina` and `fasting_blood_sugar`, and then checks for infrequent values across key categorical features (`sex`, `chest_pain_type`, `resting_ecg`, etc.). Rows containing these rare values are flagged and displayed. This helps in detecting anomalies or underrepresented categories that could affect model training or interpretation. But no categorical outliers were detected in the dataset.

## 17. Construction of Predictive models:

### Logistic Regression:

We are developing and evaluating a LR model to predict Heart\_Disease. This process involves loading the dataset, encoding ordinal variables & normalizing quantitative data for consistency. Data gets divided into train & test data. A LR model is trained on the train\_features, & results were produced on the test\_data. Finally, we evaluate this by using the model performance summary to assess performance.

```
=== Logistic Regression Results ===
Accuracy: 0.8478

Confusion Matrix:
[[68  9]
 [19 88]]

Classification Report:

```

	precision	recall	f1-score	support
0	0.78	0.88	0.83	77
1	0.91	0.82	0.86	107
accuracy			0.85	184
macro avg	0.84	0.85	0.85	184
weighted avg	0.85	0.85	0.85	184

Fig (17.1)

Logistic Regression yielded an accuracy score of **84.78%**, indicating good overall performance in forecasting Heart\_Disease. The confusion matrix exhibits that this model precisely classified 68 non-disease cases and 88 disease cases. The classification\_report underscores strong metrics for class 1 (presence of Heart\_Disease): **with precision (0.91), recall (0.82) & F1-score (0.86)**, showing it performs well in detecting heart disease. Class 0 (Negative for disease) has slightly lower precision (0.78) but high recall (0.88), meaning the model captures most healthy individuals too. Overall, the model is balanced and reliable.

Later, we implemented some techniques to improve the accuracy. Several key steps were implemented to increase the precision of the HD prediction model. To begin with, categorical features were label encoded for compatibility with the model. Then, outliers were removed using the IQR method to reduce noise and improve model stability. The dataset was then balanced using SMOTE, which synthesizes new samples for the sub-group, addressing class imbalance and helping the model learn better. We implemented feature scaling for data normalization. Finally, hyperparameter tuning was done using GridSearchCV to find the best combination of regularization strength (C) and penalty (l1, l2) for LR, leading to improved performance.

```

Best Params: {'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}
Accuracy: 0.851063829787234

Classification Report:
              precision    recall  f1-score   support

     0       0.86       0.85       0.85        71
     1       0.85       0.86       0.85        70

   accuracy          0.85
  macro avg       0.85       0.85       0.85
 weighted avg     0.85       0.85       0.85

Confusion Matrix:
[[60 11]
 [10 60]]

```

**Fig (17.2)**

We improved the LR model's accuracy from **84.78% to 85.10%** by applying key enhancements. These included removing outliers using the IQR method, balancing the dataset with SMOTE to address class imbalance, scaling features for uniformity, and fine-tuning hyperparameters using GridSearchCV. Together, these steps enhanced the model's performance and predictive reliability.

### **Random\_Forest\_Classifier:**

The workflow begins with loading the dataset, encoding categorical features using Label Encoding, numerical data we scaled to ensure balance. The data-set is then split into train & test data. To improve performance, we apply hyperparameter tuning using GridSearchCV, evaluating multiple parameter combinations—such as tree quantity, depth limits, and criteria for node splitting. This model is selected and trained. Inferences are drawn on the test data and examined using evaluation metrics and a classification-report to analyze its functionality.

```

=== Random Forest Classifier Results (Tuned) ===
Best Parameters: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}
Accuracy: 0.8804

Confusion Matrix:
[[66 11]
 [11 96]]

Classification Report:
              precision    recall  f1-score   support

     0       0.86       0.86       0.86        77
     1       0.90       0.90       0.90       107

   accuracy          0.88
  macro avg       0.88       0.88       0.88
 weighted avg     0.88       0.88       0.88

```

**Fig (17.3)**



To improve accuracy, we apply several enhancements to the Random\_Forest model. It begins with label encoding of categorical variables and removes outliers using the IQR method to reduce noise. Numerical features are standardized for consistency. It then uses Recursive Feature Elimination with Cross-Validation (RFECV) which nominates high significant attributes, reducing complexity. Finally, GridSearchCV assists in identifying the best parameter combination to find the best model configuration. These combined steps enhance the model's potential to tranform and improve prediction accuracy.

```
=== Final Random Forest Model ===
Best Params: {'class_weight': 'balanced', 'max_depth': 15, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}
Accuracy: 0.8936

Confusion Matrix:
[[61 10]
 [ 5 65]]

Classification Report:
      precision    recall  f1-score   support

     0       0.92     0.86     0.89        71
     1       0.87     0.93     0.90        70

   accuracy       0.89        141
  macro avg       0.90     0.89     0.89        141
 weighted avg       0.90     0.89     0.89        141
```

Fig (17.4)

By applying the IQR method to remove outliers and scaling the numerical features, we enhanced the standard and consistency of the training data. Outlier removal helped eliminate extreme values that could mislead the model, while feature scaling ensured all variables contributed equally. As a result, the Random Forest model became more robust and accurate, leading to an improvement in accuracy from 88.04% to 89.36%.

**XGBOOST:**

We are developing and evaluating an **XGBoost Classifier** to predict Heart\_disease. We start by loading the dataset and encoding ordinal features using Label Encoding. Data is divided onto train & test data. Without any feature scaling or outlier removal, we directly train the **XGBoost model**, a powerful and efficient gradient boosting algorithm. Once the training is over, the model's functionality is calculated by using evaluation metrics and a classification report to assess how well it predicts heart disease cases.

```
=== XGBoost Classifier Results ===
Accuracy: 0.8696

Confusion Matrix:
[[69  8]
 [16 91]]

Classification Report:
      precision    recall  f1-score   support

     0       0.81     0.90     0.85        77
     1       0.92     0.85     0.88       107

   accuracy       0.87       184
  macro avg       0.87     0.87     0.87       184
 weighted avg       0.87     0.87     0.87       184
```

Fig (17.5)

We improved the XGBoost Classifier for heart disease prediction by enhancing data preprocessing and model configuration. Ordinal features are labelled, and continuous features were normalized using StandardScaler to establish uniform scaling. We applied a stratified train-test split to maintain class balance. The XGBoost model was optimized using hyperparameters like 300 estimators and learning rate of 0.03, and max\_depth of 4, helping to reduce overfitting and boost performance. These steps together improved the model's predictive ability, which we evaluated using evaluation metrics.

```

=== XGBoost Classifier Results ===
Accuracy: 0.8967

Confusion Matrix:
[[73  9]
 [10 92]]

Classification Report:

```

	precision	recall	f1-score	support
0	0.88	0.89	0.88	82
1	0.91	0.90	0.91	102
accuracy			0.90	184
macro avg	0.90	0.90	0.90	184
weighted avg	0.90	0.90	0.90	184

**Fig (17.6)**

Initially, the XGBoost model obtained an acc. of **86.96%**. By implementing the XGBClassifier.fit() with tuned hyperparameters and applying **early stopping**, the model was better able to generalize and avoid overfitting. As a result, the accuracy improved to **89.67%**, showing the effectiveness of optimized training and regularization techniques.

### Naïve Bayes:

We build and evaluate a Gaussian Naive Bayes model to predict HD. The data-set is first added and cleaned by encoding ordinal features using LabelEncoder. The data is divided into train & test sets. A GaussianNB is fitted & evaluated on the data. The design's functionality is calculated using evaluation metrics.

```

=== Naive Bayes Classifier Results ===
Accuracy: 0.8424

Confusion Matrix:
[[65 12]
 [17 90]]

Classification Report:

```

	precision	recall	f1-score	support
0	0.79	0.84	0.82	77
1	0.88	0.84	0.86	107
accuracy			0.84	184
macro avg	0.84	0.84	0.84	184
weighted avg	0.84	0.84	0.84	184

**Fig (17.7)**

## 18. Application of SMOTE to all models:

### Accuracies of all models before applying SMOTE:

In this step, we measured the performance across all four models namely Logistic-Regression, Random\_Forest, XGBoost, & Naive-Bayes to predict HD. We started by preprocessing the dataset, including label encoding for ordinal variables and split the features into train & test sets. Each model was prepared on the same data & evaluation metrics. This allowed us to assess the merits and demerits of each model and decide which algorithm performs best for this classification task.

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	Logistic Regression	0.8696	0.8482	0.9314	0.8879	0.8957
1	Random Forest	0.8913	0.8868	0.9216	0.9038	0.9298
2	XGBoost	0.8750	0.8911	0.8824	0.8867	0.9254
3	Naive Bayes	0.8913	0.8942	0.9118	0.9029	0.9280

	Confusion Matrix
0	[[65, 17], [7, 95]]
1	[[70, 12], [8, 94]]
2	[[71, 11], [12, 90]]
3	[[71, 11], [9, 93]]

**Fig (18.1)**

1. Random\_Forest & Naive Bayes both attained the maximum acc. (89.13%), with strong recall and F1 scores (above 0.90), making them highly effective even without addressing class imbalance.
2. Naive Bayes has the highest precision (0.8942) among all, showing good ability to avoid false positives, and its ROC AUC (0.9280) is competitive with Random Forest.
3. XGBoost performs reasonably well, but its recall (0.8824) is slightly lower than the top two, making it third-best in this setting.
4. Logistic Regression shows high recall (0.9314)—meaning it identifies most positive cases—but has the lowest precision and ROC AUC, indicating more false positives and relatively weaker probability calibration.

### Observation:

Random Forest & Naive Bayes are the top contenders with balanced and high performance across all metrics. However, Random Forest slightly edges out in terms of overall AUC and F1, making it the preferred choice.

Logistic Regression, while interpretable, performs slightly worse overall, especially in terms of precision and AUC. Balancing the dataset (as you did with SMOTE later) helps refine these models further, particularly boosting Naive Bayes and reducing Logistic Regression's bias toward the majority class.

### Accuracies of all models after applying SMOTE:

In this step, **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to control category disproportion in the training data. It functions by simulating artificial samples of sub-group to stabilize the dataset. By doing this prior to model preparation, we ensure that classifiers do not become biased toward the dominant-group, leading to better generalization. This improves recall & F1 score, especially for predicting the minority group (patients with HD), making the models more stable practical medical predictions.

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	Logistic Regression	0.8587	0.8654	0.8824	0.8738	0.9034
1	Random Forest	0.9076	0.9126	0.9216	0.9171	0.9380
2	XGBoost	0.8750	0.8911	0.8824	0.8867	0.9283
3	Naive Bayes	0.8967	0.8952	0.9216	0.9082	0.9316

	Confusion Matrix
0	[[68, 14], [12, 90]]
1	[[73, 9], [8, 94]]
2	[[71, 11], [12, 90]]
3	[[71, 11], [8, 94]]

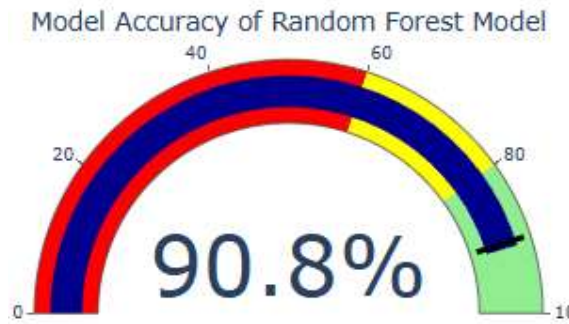
**Fig (18.2)**

1. Random-Forest surpasses all other models with greatest accuracy (90.76%), F1 Score (91.71%), and ROC AUC (0.9380). It balances both precision and recall, making it the best-performing model overall after applying SMOTE.
2. Naive Bayes significantly improves after SMOTE. The F1 score (90.82%) and ROC AUC (0.9316) are second only to Random Forest, indicating strong generalization on the balanced data.
3. XGBoost also performs well but ranks third. While it retains a high ROC AUC (0.9283), its F1 score (88.67%) is lower than the top two.
4. Logistic Regression remains the least performing model post-SMOTE, although it still maintains decent metrics (F1 Score: 87.38%, ROC AUC: 0.9034). It's a good interpretable baseline, but lags in overall accuracy.

### Observation:

After handling class imbalance using SMOTE, all models show noticeable improvement. However, Random Forest emerges as the top performer, offering the best trade-off between accuracy, precision, recall, and AUC. Naive Bayes also becomes a surprisingly competitive model. While XGBoost and Logistic Regression are still effective, they are slightly behind in precision and F1 performance. Thus, the RF model is an advisable model to determine heart-disease.

## 19. Evaluation:



**Fig (19.1)**

The RF model attains a top result of all four models. Random Forest often achieves the **highest accuracy** in classification tasks like heart disease prediction for several reasons:

1. **Ensemble Learning:** Several decision\_trees are created & the results are summed up to reduce overfitting issues and to improve the transferability compared to a single tree.
2. **Handles Feature Interactions Automatically:** Random Forest naturally captures non-linear realtionships between features, such as the relationship between cholesterol, age, and chest pain.
3. **Robust to Noise and Outliers:** Since it averages results from many trees, Random Forest is more tolerant to outliers & noisy data, making it reliable for real-world health datasets.
4. **Feature Randomization:** By selecting a random subset of features for every division, it promotes variety among trees and diminishes correlation, increasing model stability and accuracy.
5. **Less Assumption on Data Distribution:** Unlike models like Naive Bayes or Logistic Regression, Random Forest doesn't assume any distribution or linearity, making it more flexible for complex, non-linear data.

Random Forest achieved the highest accuracy because it combines multiple decision trees, handles complex patterns well, and is less prone to overfitting, especially after balancing the data with SMOTE.

**ROC Interpretation for Random\_Forest\_Classifier:**

The ROC for the Random\_Forest shows excellent classification performance, with an **AUC of 0.93**. The output shows that the model possesses a **high TP value** while maintaining a **low FP value**, effectively differentiating between patients' presence/absence of Heart\_Disease. Area\_Under\_Curve values closer to 1.0 depict better performance, and 0.93 reflects that the Random\_Forest is very efficient at predicting HD.

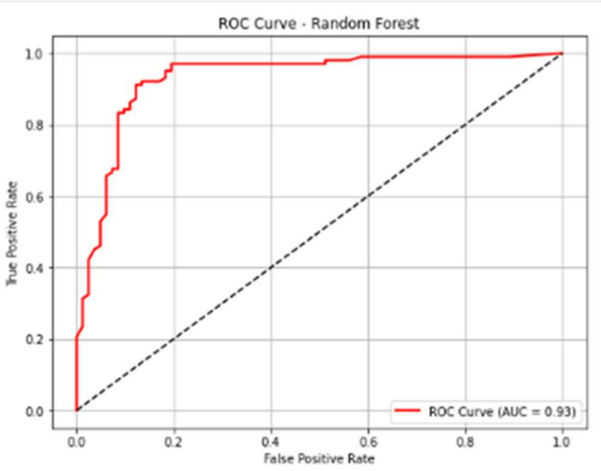


Fig (19.2)

**Model Explainability - LIME - Local Interpretable Model-Agnostic Explanations:**

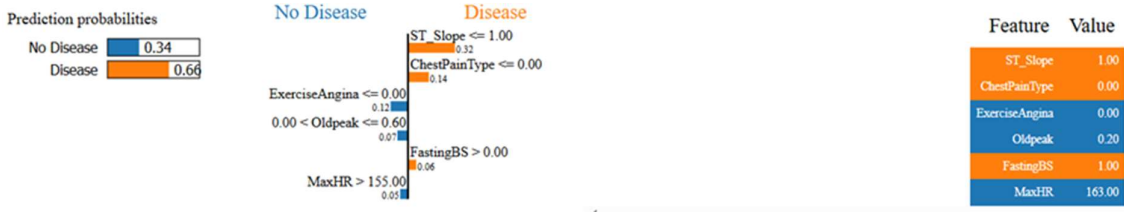


Fig (19.3)

**Prediction Summary:**

The model predicts the person has **heart disease** with **66% probability**, compared to 34% for no disease.

**Features Indicating Disease:**

- **ST\_Slope ≤ 1.0** had the biggest impact (+0.32), indicating abnormal heart behavior during stress.
- **ChestPainType = 0** (likely asymptomatic) also increased the risk.
- **FastingBS > 0** (high blood sugar) slightly contributed to the disease prediction.

● **Features Indicating No Disease:**

- **No exercise-induced angina** reduced the risk (−0.12).
- **Oldpeak = 0.2**, a mild ST depression, had a small protective effect.
- **MaxHR = 163**, a good heart rate, helped slightly lower the risk.

Features Supporting "Disease" (Orange Bars)

Feature	Contribution	Explanation
ST_Slope <= 1.0	+0.33	A flat ST slope is strongly associated with disease.
ChestPainType <= 0.0	+0.14	Encoded type 0 may be typical angina — likely indicating heart issues.
FastingBS > 0.0	+0.07	High fasting blood sugar (i.e., diabetes risk) contributes to disease likelihood.

Features Supporting "No Disease" (Blue Bars)

Feature	Contribution	Explanation
ExerciseAngina <= 0.0	-0.13	No exercise-induced angina reduces disease likelihood.
0.00 < Oldpeak <= 0.6	-0.09	Low ST depression suggests less heart strain — healthy sign.
MaxHR > 155.0	-0.06	High maximum heart rate usually reflects good fitness levels.

Feature Values (Bottom Table)

Feature	Value	Meaning
ST_Slope	1.00	Flat slope (possibly unhealthy)
ChestPainType	0.00	Encoded pain type (e.g., typical angina)
ExerciseAngina	0.00	No angina during exercise
Oldpeak	0.20	Very low ST depression — good
FastingBS	1.00	Fasting blood sugar is high (risk)
MaxHR	163.00	Very high HR — typically a healthy sign

**Fig (19.4)**

**Final Insight:**

The model predicts heart disease for this patient, primarily due to:

- A flat or abnormal ST slope
- Asymptomatic chest pain
- High blood sugar

These outweigh the healthier indicators like no exercise-angina and a high maximum heart rate. Despite some healthy indicators, strong risk factors like abnormal ST slope and asymptomatic chest pain led the model to predict heart disease.

This type of LIME explanation helps in comprehending the rationale behind a black-box model's specific choice and aiding clinical decision-making.

## 20. Model Deployment-Interactive Dashboard:

I created an interactive web\_app with Streamlit that enables users to estimate the probability of HD derived from input characteristics viz. Age\_group,type of chest\_pain, lipid levels, ST\_slope, and additional factors. The app takes user inputs, preprocesses them, and passes them through a trained ML model (like Random\_Forest or XGBoost). It then displays the prediction along with the probability of having heart disease. Additionally, the app includes a LIME-based explanation panel that highlights which features contributed most to the prediction, making it easy for users and clinicians to understand model decisions transparently.

**Step 1:** Click on this Pink button in Jupyter to open the app. It takes you to a new tab.



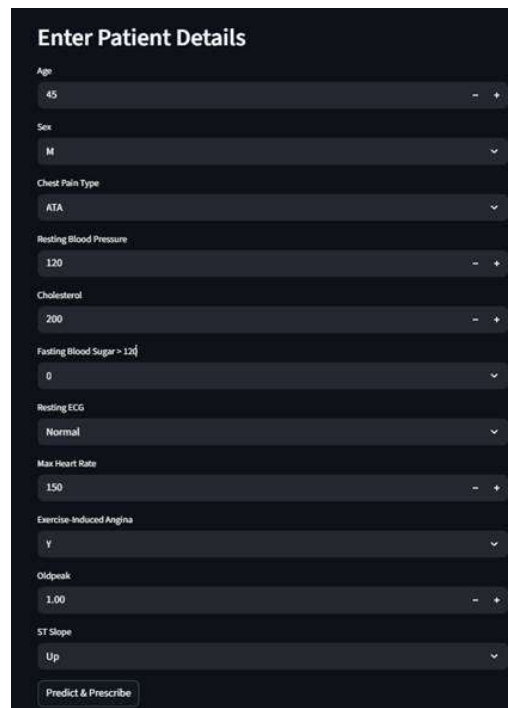
**Fig (20.1)**

**Step 2:** Wake up the app and wait for 1 min while the page says “Your app is in the oven.”



**Fig (20.2)**

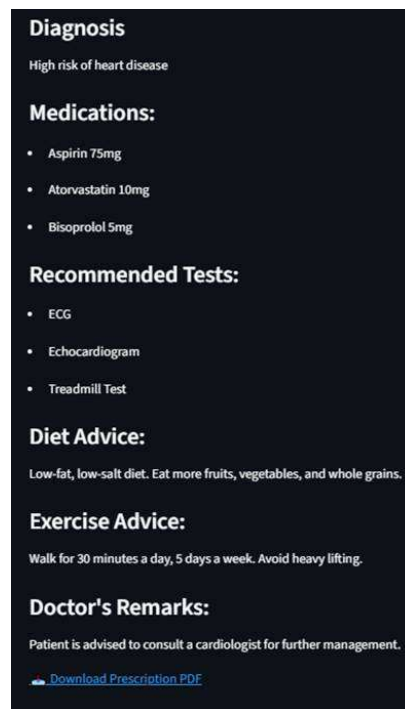
**Step 3:** The app shows this page where we have to enter the details of the patients.

A dark-themed web form titled "Enter Patient Details" in white text. The form contains several input fields with labels in white: "Age" (numeric, value 45), "Sex" (dropdown, value M), "Chest Pain Type" (dropdown, value ATA), "Resting Blood Pressure" (numeric, value 120), "Cholesterol" (numeric, value 200), "Fasting Blood Sugar > 126" (dropdown, value 0), "Resting ECG" (dropdown, value Normal), "Max Heart Rate" (numeric, value 150), "Exercise-Induced Angina" (dropdown, value Y), "Oldpeak" (numeric, value 1.00), and "ST Slope" (dropdown, value Up). At the bottom of the form is a button labeled "Predict & Prescribe" in white text.

**Fig (20.3)**



This image shows the results of the diagnosis of a patient with a high risk of cardiac arrest.



**Fig (20.4)**

The HD prediction app provides a fast, convenient, and accessible application for initial risk assessment. Leveraging machine learning models helps identify potential heart disease risks based on clinical inputs, supporting proactive medical attention. The use of explainable AI techniques (like LIME) enhances transparency, making predictions more trustworthy for both medical professionals and patients. Overall, this app serves as a valuable decision-support tool to assist in early diagnosis, reduce manual workload, and promote preventive care in healthcare settings.

## **21. Limitations:**

While the heart disease prediction project demonstrates promising results using AI techniques, several limitations should be acknowledged:

### **1. Small Dataset Size**

The project is based on a relatively small dataset, which reduces the applicability of the results to other contexts. With only a few hundred records, the models may operate successfully on the train\_data but fail to extend effectively on future input populations.

### **2. Model Overfitting Risk**

Due to the limited data, complex models (e.g., Random Forest, XGBoost) may overfit, capturing noise instead of meaningful patterns. Though validation techniques like cross-validation were used, overfitting remains a concern.

3. **Lack of Real-Time Validation**

The model was not validated on real-world or live clinical data. All testing was done on historical data from a single dataset, which may not represent real-time patient variability or diverse clinical settings.

4. **Simplified Feature Set**

The project only used the features available in the dataset, which primarily include physiological and clinical metrics. Important lifestyle and behavioral variables such as diet, stress, exercise, and family history were not considered, limiting the model's comprehensiveness.

5. **Interpretability Challenges**

While models like Random Forest and XGBoost offer good results they are harder to interpret compared to models like logistic regression. Though LIME was used to address interpretability, more extensive explainability tools are required for deployment in clinical practice.

6. **Static, Cross-Sectional Data**

The dataset used is static and cross-sectional, meaning it captures one snapshot promptly. This controls the capacity to model disease progression/ predict future outcomes, which are critical in clinical decision-making.

7. **Lack of Clinical Integration**

The model was not tested in a clinical environment. As such, it lacks integration with electronic health records (EHRs) or decision support systems, which would be necessary for real-world medical use.

8. **Ethical and Privacy Considerations**

The project did not involve real patients, so data privacy and ethics issues were minimal. However, deploying such models in real-world settings would require strict compliance with healthcare regulations (e.g., HIPAA, GDPR) to ensure patient confidentiality and ethical use of AI.

## **22. Future Works:**

Building upon the current heart disease prediction project, several directions can be explored to enhance its operational success, trustworthiness, and applicability in real environments:

1. **Expand Dataset Size and Diversity**

Future work must incorporate the integration of bigger & varied datasets sourced from multiple hospitals and areas. This will enhance the model's generalizability and reduce overfitting while ensuring representation across different demographics and clinical backgrounds.

2. **Include Additional Risk Factors**

Incorporating lifestyle and behavioral features such as smoking, exercise, diet, alcohol consumption, and family history can significantly improve model accuracy and clinical relevance. These variables are known contributors to cardiovascular risk and should be collected in future datasets.

### **3. Deploy Time-Series and Longitudinal Models**

Using time-series data to track patient health over time would allow the creation of forecasting models that forecast illness advancement & treatment response. This could provide valuable insights for early intervention and long-term patient monitoring.

### **4. Real-World Clinical Validation**

The models should be tested and validated in collaboration with healthcare professionals in real clinical settings. This would allow for feedback from practitioners, helping to assess the feasibility and utility of the system in everyday medical decision-making.

### **5. Integration with Electronic Health Records (EHRs)**

Developing systems that integrate seamlessly with existing EHR platforms will allow for automatic data ingestion and real-time prediction. This will increase the efficiency of clinical workflows and improve early detection.

### **6. Ethical, Legal, and Regulatory Compliance**

Future deployments should prioritize data privacy, security, and ethical considerations by adhering to regulations such as HIPAA, GDPR, and local healthcare data laws. Accurate documentation, consent processes, and equity evaluations need to be established.

## **23. Conclusion:**

This project intended to develop a prognostic model for HD using ML methods applied to structured clinical data. By leveraging frameworks viz. Logistic Regression, Random Forest, Naive Bayes, and XGBoost, along with interpretability tools like LIME, the project successfully demonstrated the potential of data-driven approaches in supporting early detection and evaluation of Heart\_Disease. "As part of preprocessing, one-hot coding was applied to cast the ordinal variables to quantitative type, followed by attribute filtering were applied to make sure the grade and significance of the predictor variables. The models were assessed using standard performance standards, and among them, hybrid methods such as Random\_Forest and XGBoost provided the most accurate and robust results.

While the results are promising, the project is not without limitations, such as a small dataset size, lack of behavioral features, and absence of real-time clinical testing. Nevertheless, the project yields a strong base for subsequent study and clinical integration. With further development—including extensive data, practical proof, and better clarity—the model can be very useful in preventive cardiology, enabling healthcare professionals to contribute to well-informed medical decisions and better recovery rates. To conclude, this project underscores the effectiveness of AI in medical diagnostics and sets the stage for more advanced, ethical, and patient-centered AI applications in healthcare.

## 24. References:

1. Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics, Advance online publication. <https://doi.org/10.1016/S2772442522000016>
2. Hossain, M. I., Maruf, M. H., Khan, M. A. R., Prity, F. S., Fatema, S., Ejaz, M. S., & Khan, M. A. S. (2023). Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison. Iran Journal of Computer Science, 6(4), 397–417. <https://doi.org/10.1007/s42044-023-00148-7>  
researchgate.net+5link.springer.com+5scribd.com+5
3. El-Sofany, H., Bouallegue, B., & Abd El-Latif, Y.(M.(A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14(1), 23277. <https://doi.org/10.1038/s41598-024-74656-2>
4. Alshraideh, M. A., Alshraideh, N., Alshraideh, A., Alkayed, Y., Al Trabsheh, Y., & Alshraideh, B. (2024). Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital. Applied Computational Intelligence and Soft Computing, 2024, Article 5080332. <https://doi.org/10.1155/2024/5080332>  
(pmc.ncbi.nlm.nih.gov)
5. Islam, M. A. (2024). Precision healthcare: A deep dive into machine learning algorithms and feature selection techniques for accurate heart disease prediction [Article]. [Journal Name]. Advance online publication. <https://doi.org/S001048252400516X> sciencedirect.com