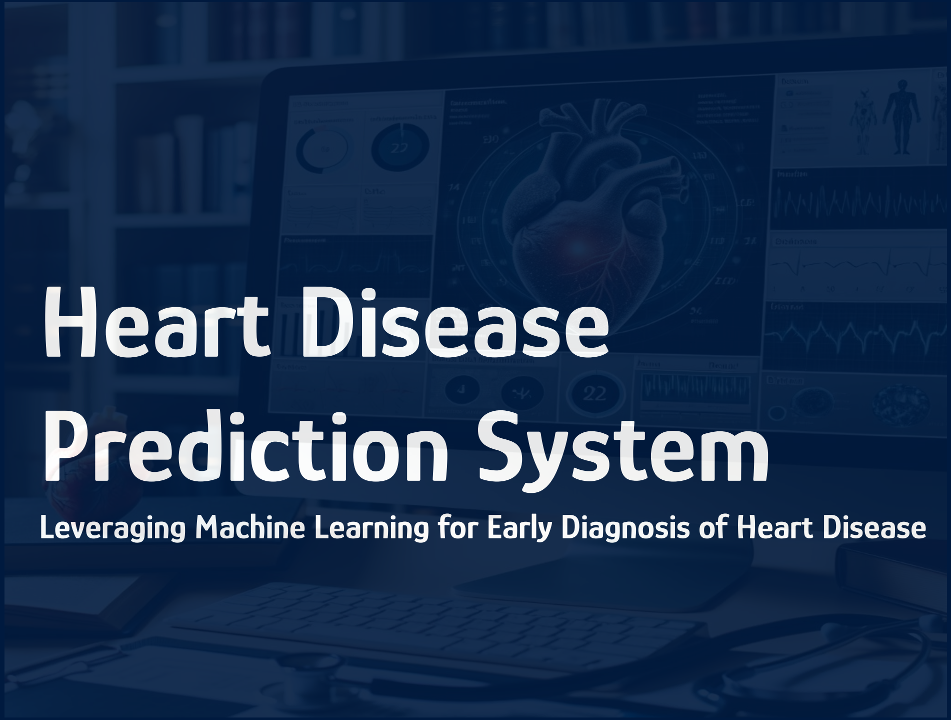
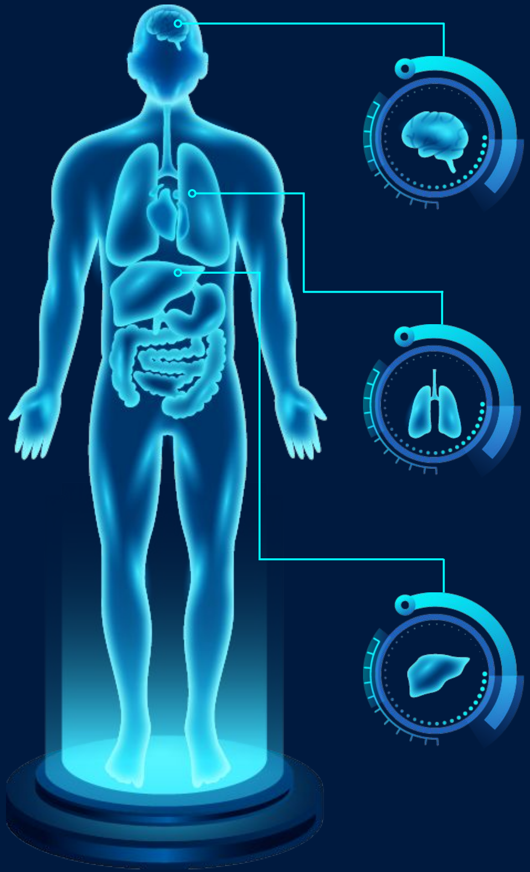


# Heart Disease Prediction System

## Leveraging Machine Learning for Early Diagnosis of Heart Disease

MEF UNIVERSITY
Your Freedom in Learning

Batuhan Birinci 042001047
Meryem Rana Türker 042001021



# Contents

01

Introduction

02

Dataset

03

Methodology

04

Interface

05

Insights

06

Results

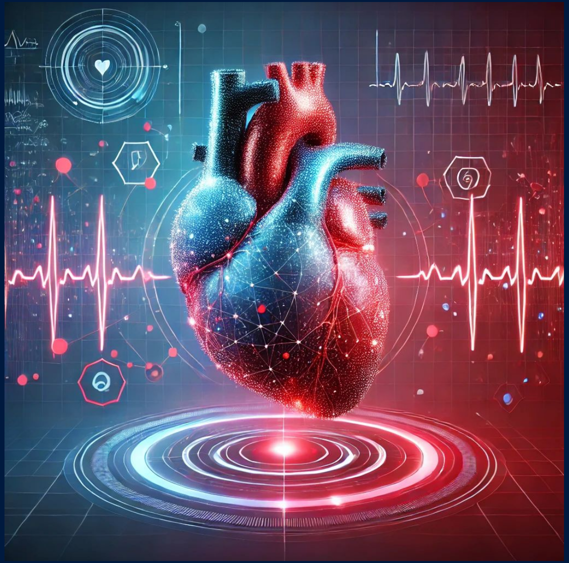
# Introduction

Cardiovascular diseases cause approximately 17.9 million deaths annually, surpassing most cancer mortality rates. Early survival prediction and risk factor identification are critical for timely interventions.

Recent advancements in computer science, particularly in machine learning and artificial intelligence, have transformed medical research, enabling more accurate diagnostics and predictions.

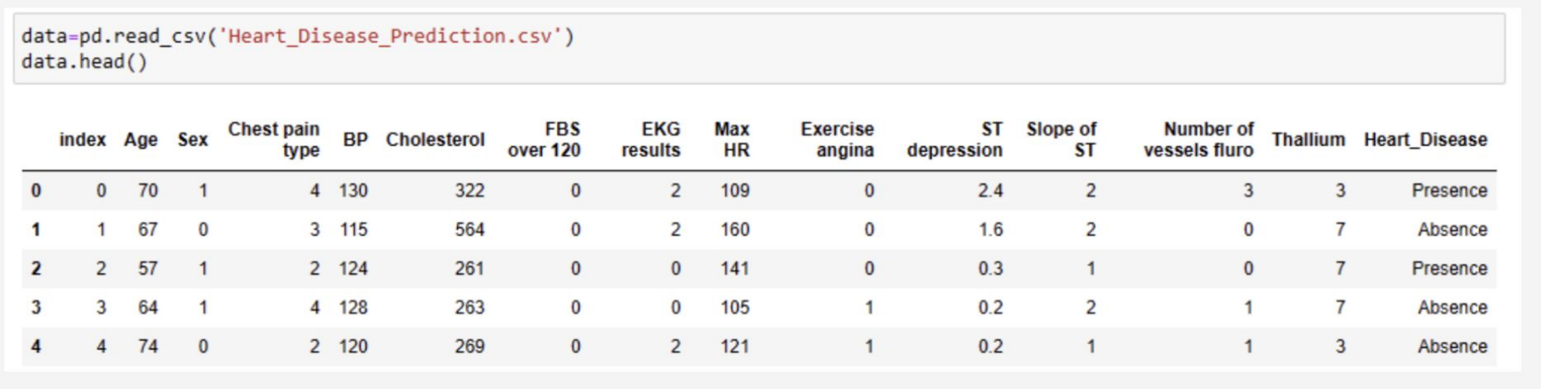
This project applies data analysis and visualization techniques to predict heart disease risks using a publicly available dataset. A web-based tool, developed with Flask and Python, provides an interactive interface for risk assessment.

**Keywords**: Heart Disease Prediction, Data Analysis, Medical AI, Machine Learning



# Dataset Overview

- **Source**: UCI Heart Disease Dataset
- **Data Attributes**:
  - **Features**: Age, Sex, Resting Blood Pressure, Cholesterol, etc. (13 key features used for prediction).
  - **Target Variable**: Presence (1) or absence (0) of heart disease.
- **Exploratory Data Analysis (EDA)**:
  - Distribution of features (e.g., histogram of Age, Cholesterol levels).
  - Correlation matrix highlighting relationships among features.



```
data=pd.read_csv('Heart_Disease_Prediction.csv')
data.head()
```

| | index | Age | Sex | Chest pain type | BP | Cholesterol | FBS over 120 | EKG results | Max HR | Exercise angina | ST depression | Slope of ST | Number of vessels fluro | Thallium | Heart_Disease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 70 | 1 | 4 | 130 | 322 | 0 | 2 | 109 | 0 | 2.4 | 2 | 3 | 3 | Presence |
| **1** | 1 | 67 | 0 | 3 | 115 | 564 | 0 | 2 | 160 | 0 | 1.6 | 2 | 0 | 7 | Absence |
| **2** | 2 | 57 | 1 | 2 | 124 | 261 | 0 | 0 | 141 | 0 | 0.3 | 1 | 0 | 7 | Presence |
| **3** | 3 | 64 | 1 | 4 | 128 | 263 | 0 | 0 | 105 | 1 | 0.2 | 2 | 1 | 7 | Absence |
| **4** | 4 | 74 | 0 | 2 | 120 | 269 | 0 | 2 | 121 | 1 | 0.2 | 1 | 1 | 3 | Absence |

# Methodology

## Data Preparation

The project utilized the UCI Heart Disease Dataset, which includes 13 features (e.g., Age, Cholesterol, Blood Pressure) and a binary target. Data preprocessing involved Min-Max Scaling for normalization, handling missing values, and balancing the dataset using SMOTE. Key features were selected through correlation analysis and domain expertise to enhance model efficiency.

## Model Development

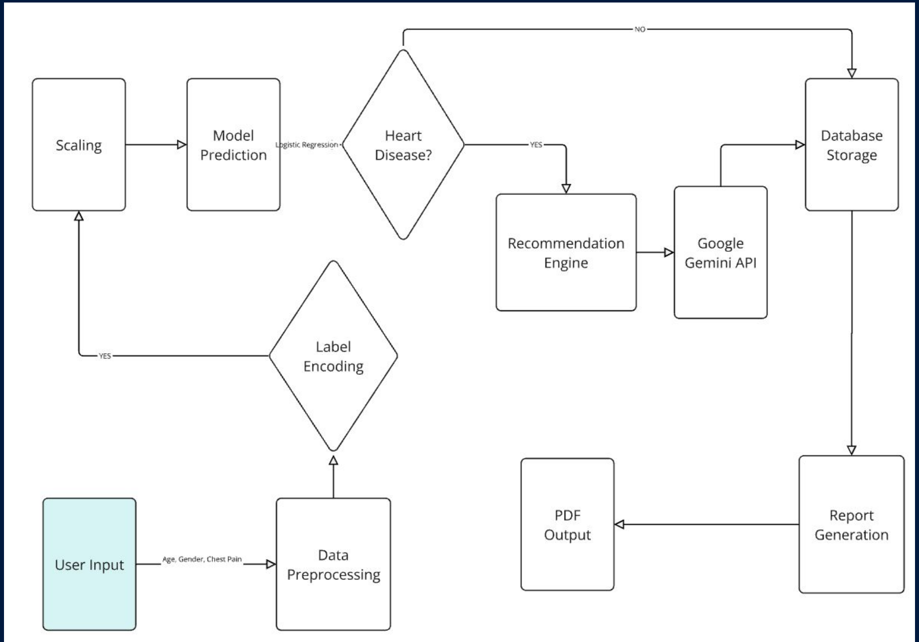
A Logistic Regression model was trained on 80% of the dataset and tested on the remaining 20%. The model achieved an **accuracy of 90.7%** and an **ROC-AUC score of 94.8%**, demonstrating its capability to distinguish between the presence and absence of heart disease effectively.

## Web Application Integration

The trained model was deployed using Flask to create a web application with an interactive Bootstrap-based interface. Users input their data, receive risk classifications ("Low" or "High") displayed with a visual risk bar, and access personalized recommendations. Additional features include QR code generation for sharing results.

# System Architecture

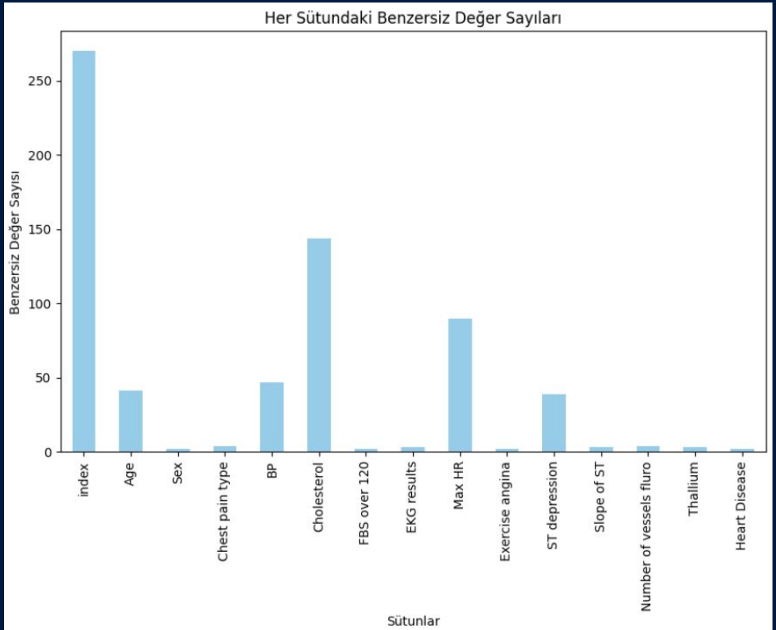
- **Data Input:** User provides patient data (e.g., age, gender, chest pain) via a web interface.
- **Preprocessing:** Data normalized with Standard Scaler and encoded using Label Encoding.
- **Model Prediction:** Logistic Regression predicts heart disease risk with confidence scores (e.g., "High Risk").
- **Recommendation Engine:** Google Gemini API offers personalized recommendations based on predictions.
- **Database:** SQLite stores input data, predictions, and confidence scores.
- **PDF Report:** Generates downloadable reports summarizing predictions and insights.
- **User Interface:** Flask app with routes for predictions and report downloads.



# DETERMINATION OF CATEGORICAL AND NUMERICAL VARIABLES

In this chart, columns are separated based on the number of unique values: columns with a high number of unique values typically contain numerical data (e.g., age or cholesterol), while columns with a low number of unique values usually represent categorical data (e.g., gender or ECG results).

This approach has been used to distinguish categorical and numerical variables.

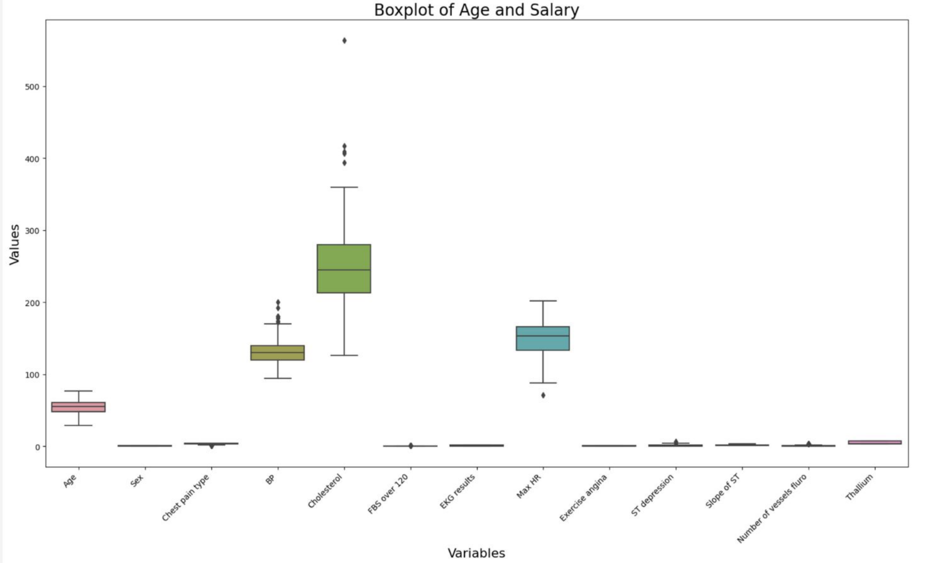


## CATEGORICAL VARIABLES:

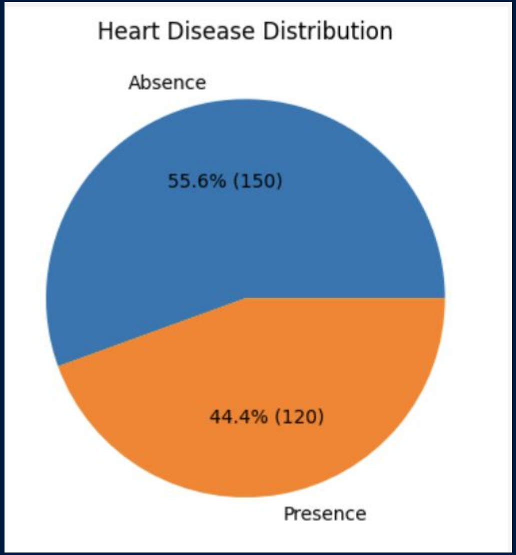
- **Sex:** Gender
- **Chest pain type:** Type of chest pain
- **FBS over 120:** Whether fasting blood sugar is over 120
- **EKG Results:** Results of the electrocardiogram
- **Exercise Angina:** Presence of exercise-induced angina
- **Slope of ST:** Slope of the ST segment
- **Number of vessels fluoroscopy:** Number of vessels detected in fluoroscopy
- **Thallium:** Thallium scan results

## NUMERICAL VARIABLES:

- **Age:** Age
- **BP:** Blood pressure
- **Cholesterol:** Cholesterol level
- **Max HR:** Maximum heart rate
- **ST Depression:** ST depression level



# TARGET VARIABLE ANALYSIS



**Target Variable:** Presence (1) or Absence (0) of heart disease.

**From this chart, it can be inferred that the target variable has a balanced distribution.**
The target, also referred to as the label column (Heart Disease), is divided into two categories: **Absence** and **Presence.**

- **Absence:** Indicates no disease, meaning the individual is healthy.
- **Presence:** Indicates the presence of the disease, meaning the individual is unhealthy.

The distribution of these categories is illustrated in a pie chart.

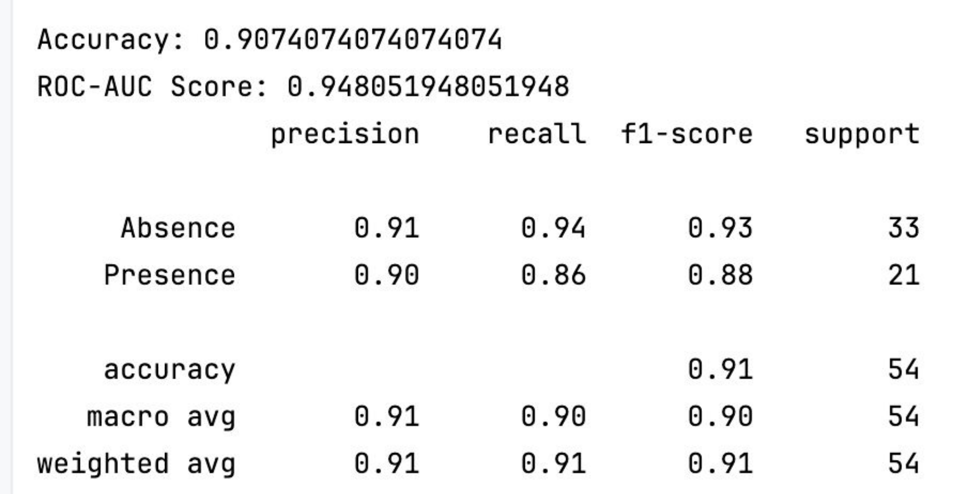
# Classifier: Logistic Regression

| Parameter | Description |
|---|---|
| **Algorithm** | Logistic Regression |
| **Target Variable** | Heart Disease (0 = Absence, 1 = Presence) |
| **Independent Variables** | All columns except `Heart Disease` |
| **Data Split** | 80% Training, 20% Testing |
| **Maximum Iterations (max_iter)** | 1000 |
| **Evaluation Metrics** | Accuracy, ROC-AUC Score, Classification Report |
| **Prediction Probability** | Calculated for the positive class (Heart Disease = 1). |
| **Risk Levels** | High, Medium, Low (determined by threshold values). |

| Risk Level | Threshold | Formula |
|---|---|---|
| **High Risk** | $\text{Probability} \geq 0.7$ | `Risk = Probability × 100` |
| **Medium Risk** | $0.4 \leq \text{Probability} < 0.7$ | `Risk = Probability × 50 + 50` |
| **Low Risk** | $\text{Probability} < 0.4$ | `Risk = Probability × 30` |

| File | Description |
|---|---|
| `prediction_results.csv` | Contains actual labels, predictions, probabilities, and risk rates. |
| `heart_disease_model.pkl` | Serialized logistic regression model. |

# Classification Parameters



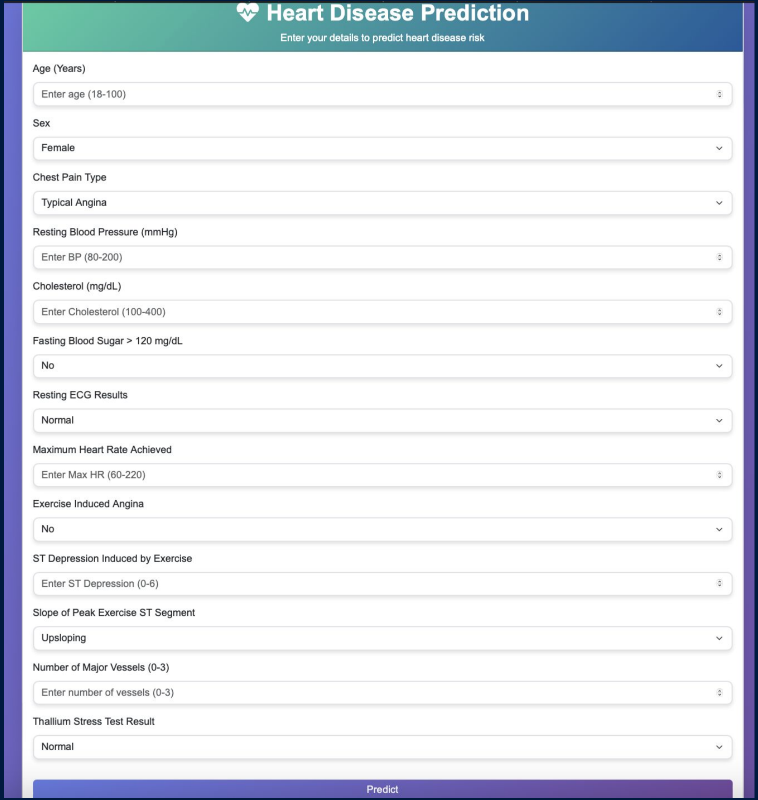
```
Accuracy: 0.9074074074074074
ROC-AUC Score: 0.948051948051948
              precision    recall  f1-score   support

     Absence       0.91      0.94      0.93        33
    Presence       0.90      0.86      0.88        21

    accuracy                           0.91        54
   macro avg       0.91      0.90      0.90        54
weighted avg       0.91      0.91      0.91        54
```

The Classification Report includes metrics such as precision, recall, F1-score, and support for the model's two classes ('Absence' and 'Presence').
Additionally, the report provides the model's overall accuracy (91%), macro average, and weighted average values.
This report allows us to evaluate in detail how well the model performs for each class.

# Project Interface



## Heart Disease Prediction

Enter your details to predict heart disease risk

Age (Years)

Enter age (18-100)

Sex

Female

Chest Pain Type

Typical Angina

Resting Blood Pressure (mmHg)

Enter BP (80-200)

Cholesterol (mg/dL)

Enter Cholesterol (100-400)

Fasting Blood Sugar > 120 mg/dL

No

Resting ECG Results

Normal

Maximum Heart Rate Achieved

Enter Max HR (60-220)

Exercise Induced Angina

No

ST Depression Induced by Exercise

Enter ST Depression (0-6)

Slope of Peak Exercise ST Segment

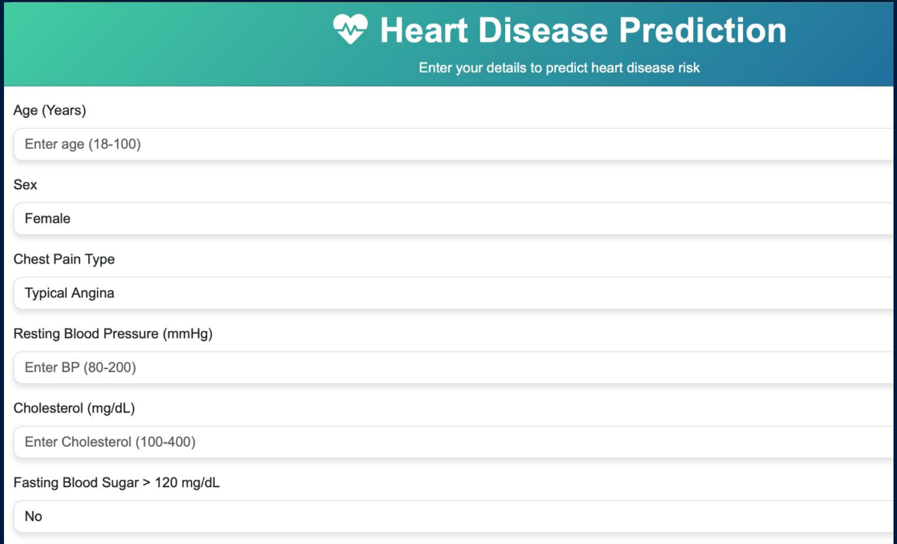
Upsloping

Number of Major Vessels (0-3)

Enter number of vessels (0-3)

Thallium Stress Test Result

Normal



Heart Disease Prediction

Enter your details to predict heart disease risk

Age (Years)

Enter age (18-100)

Sex

Female

Chest Pain Type

Typical Angina

Resting Blood Pressure (mmHg)

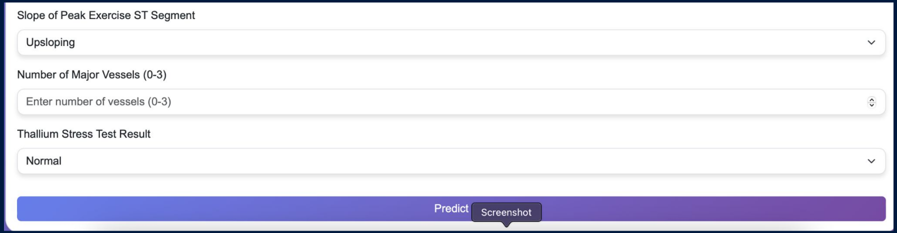
Enter BP (80-200)

Cholesterol (mg/dL)

Enter Cholesterol (100-400)

Fasting Blood Sugar > 120 mg/dL

No



Slope of Peak Exercise ST Segment

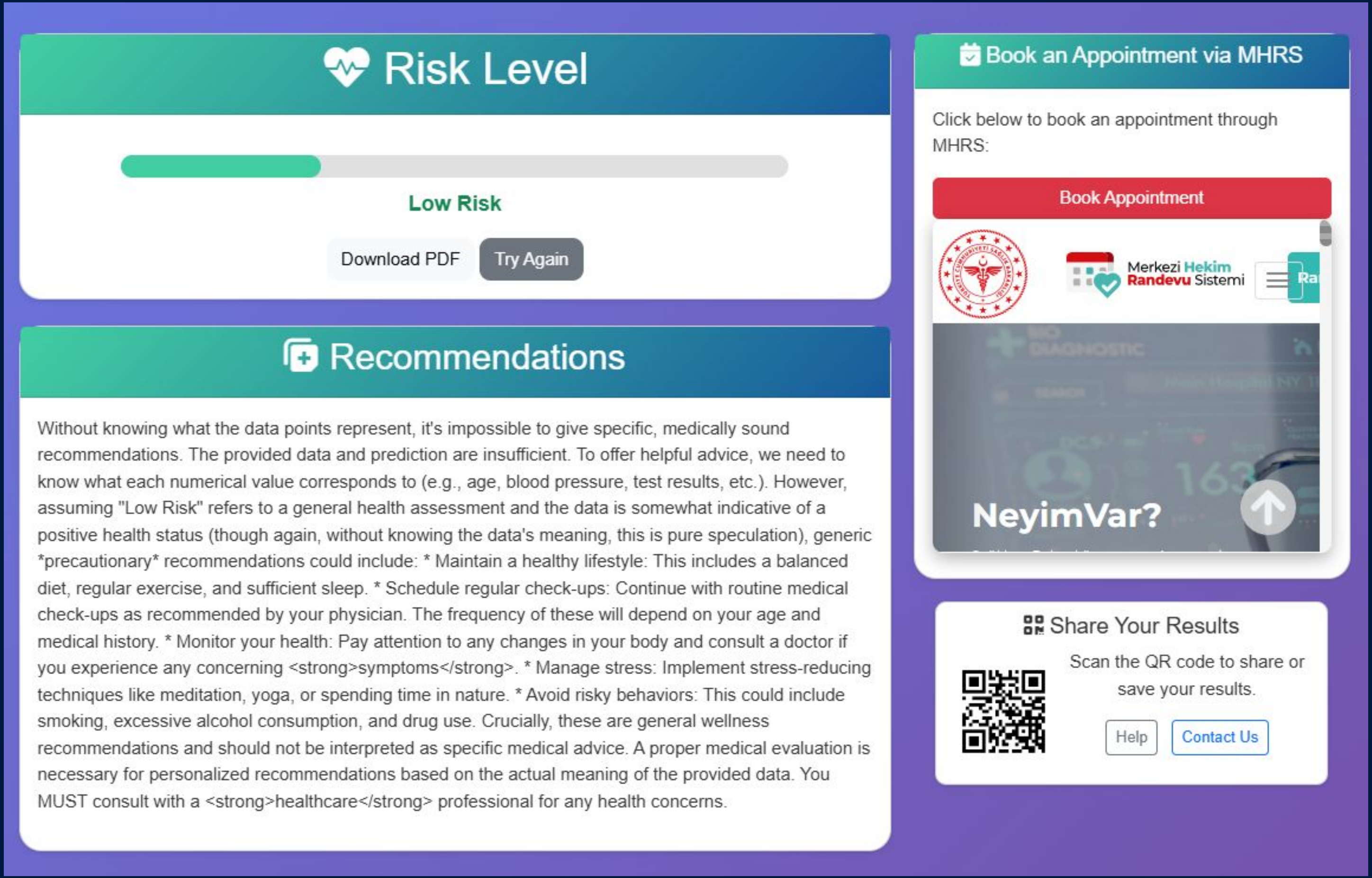
Upsloping

Number of Major Vessels (0-3)

Enter number of vessels (0-3)

Thallium Stress Test Result

Normal



## Risk Level

Low Risk

## Recommendations

Without knowing what the data points represent, it's impossible to give specific, medically sound recommendations. The provided data and prediction are insufficient. To offer helpful advice, we need to know what each numerical value corresponds to (e.g., age, blood pressure, test results, etc.). However, assuming "Low Risk" refers to a general health assessment and the data is somewhat indicative of a positive health status (though again, without knowing the data's meaning, this is pure speculation), generic *precautionary* recommendations could include: * Maintain a healthy lifestyle: This includes a balanced diet, regular exercise, and sufficient sleep. * Schedule regular check-ups: Continue with routine medical check-ups as recommended by your physician. The frequency of these will depend on your age and medical history. * Monitor your health: Pay attention to any changes in your body and consult a doctor if you experience any concerning <strong>symptoms</strong>. * Manage stress: Implement stress-reducing techniques like meditation, yoga, or spending time in nature. * Avoid risky behaviors: This could include smoking, excessive alcohol consumption, and drug use. Crucially, these are general wellness recommendations and should not be interpreted as specific medical advice. A proper medical evaluation is necessary for personalized recommendations based on the actual meaning of the provided data. You MUST consult with a <strong>healthcare</strong> professional for any health concerns.

### Book an Appointment via MHRS

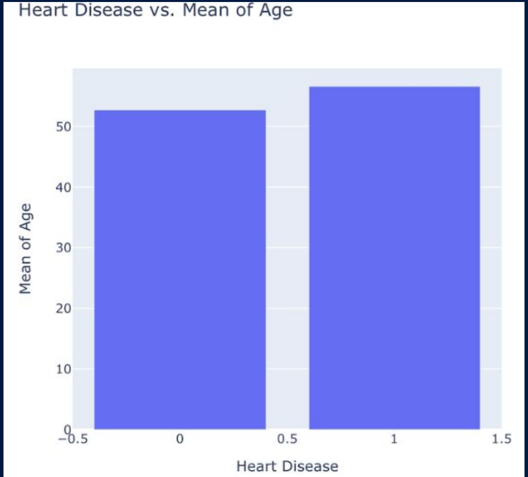
Click below to book an appointment through MHRS:

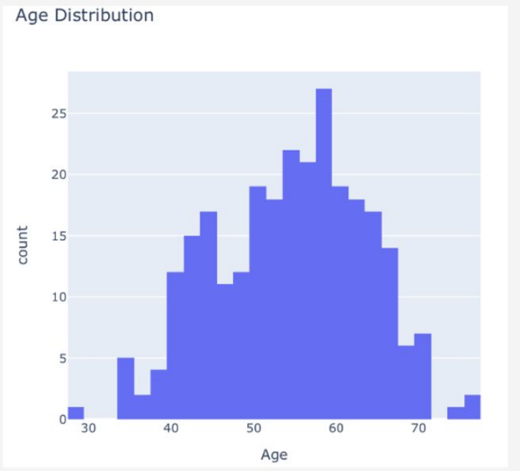
### Share Your Results

Scan the QR code to share or save your results.

# Insights

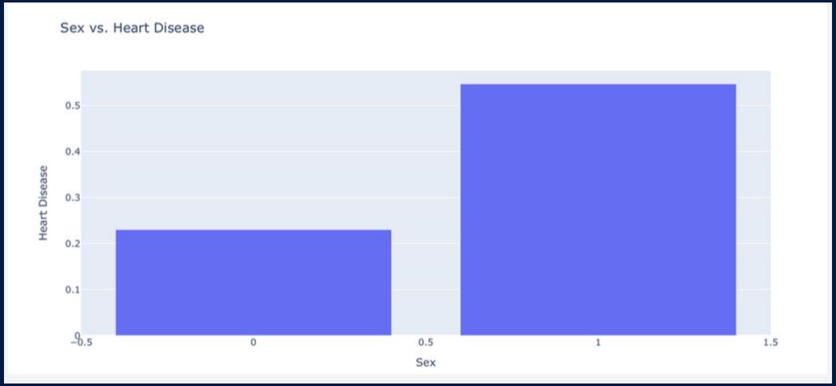
1. **AGE**
    - **Relationship Between Age and Heart Disease:**
      The average age of individuals with heart disease is higher than those without it. This indicates that age is a significant risk factor for heart disease.
    - **Risk Groups:**
      The higher average age of individuals with heart disease suggests that the risk increases with age. Older individuals are at a higher risk.
    - **Average Age:**
        - Individuals without heart disease (0): ~52-53 years.
        - Individuals with heart disease (1): ~56-57 years.
    - **Most Common Age Group:**
      The most prevalent age group in the data is 55-60, which also corresponds to a higher frequency of heart disease.

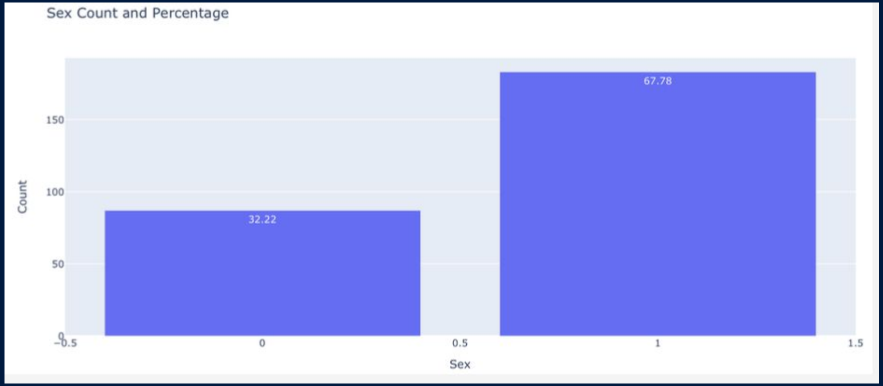




## SEX

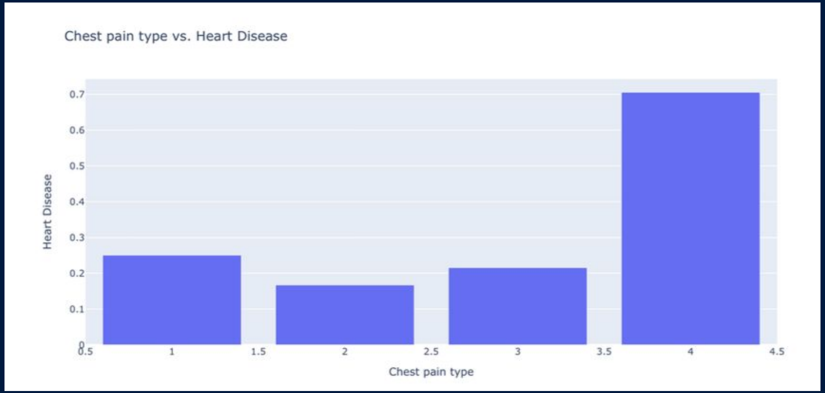
- **Heart Disease Prevalence:**
  - Women (0): ~20% (or 0.2).
  - Men (1): ~50% (or 0.5).
  - Men have a significantly higher prevalence of heart disease compared to women.

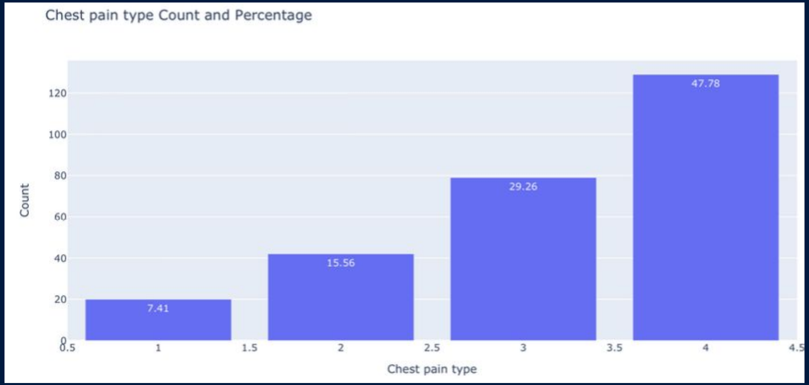




## CHEST PAIN TYPE

- **Risk Levels by Chest Pain Type:**
  - Type 4 chest pain is associated with a very high risk of heart disease and requires immediate medical attention.
  - Type 2 chest pain has the lowest risk of heart disease, suggesting it carries less danger compared to other types.

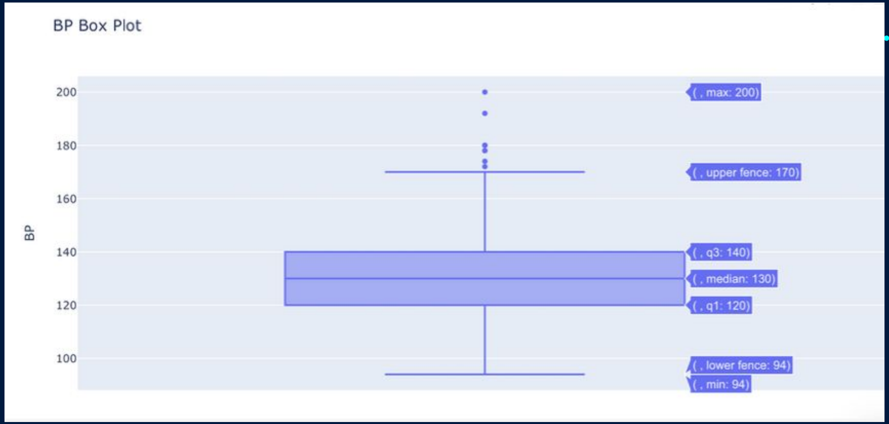




  - Type 4 chest pain is both common and the most dangerous in terms of heart disease risk.
  - Type 2 chest pain is less common and the least dangerous.

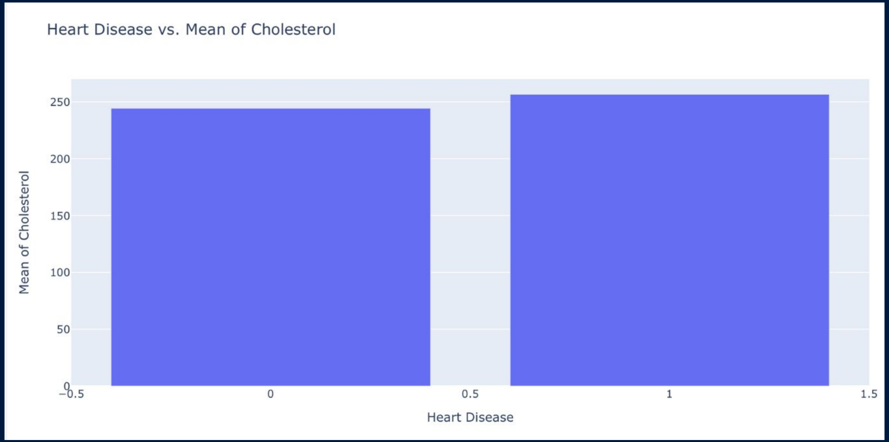
## BP (Blood Pressure)

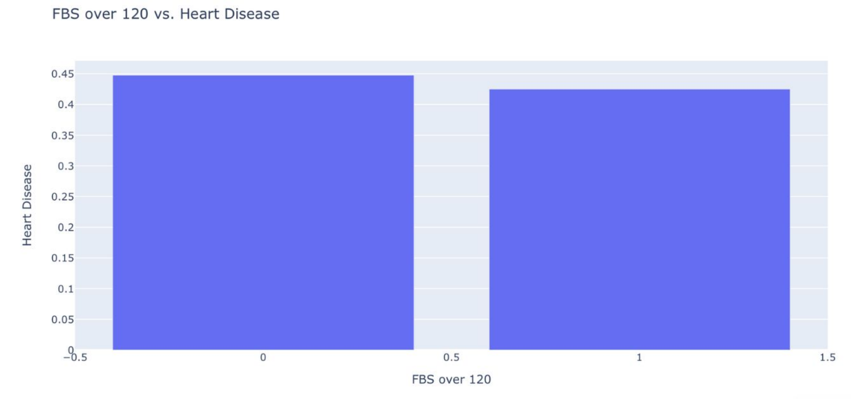
- Average blood pressure values (~130) are similar for individuals with and without heart disease, indicating no significant difference.



## CHOLESTEROL

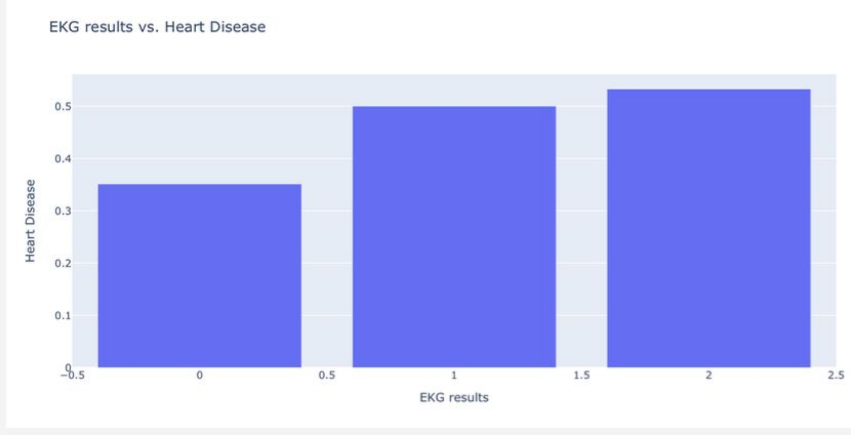
- Average cholesterol levels (~240) are almost identical for both groups, suggesting no strong correlation between cholesterol and heart disease presence.





## FBS (Fasting Blood Sugar)

- **FBS ≤ 120 (0):** Heart disease prevalence ~45%.
- **FBS > 120 (1):** Heart disease prevalence ~40%.
- FBS values above or below 120 show no significant difference in heart disease prevalence, which ranges between 40%-45%.

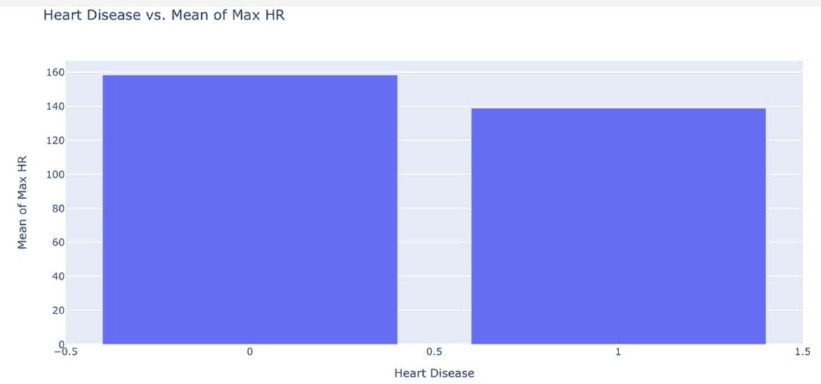


## EKG (Electrocardiogram Results)

- **Heart Disease Prevalence by EKG Results:**
  - EKG Result = 1 or 2: ~50%.
  - EKG Result = 0: ~30%.
- **Observations:**
  - EKG Result 0 is mostly normal.
  - Results 1 and 2 indicate higher risk levels, especially result 2 showing specific abnormalities.

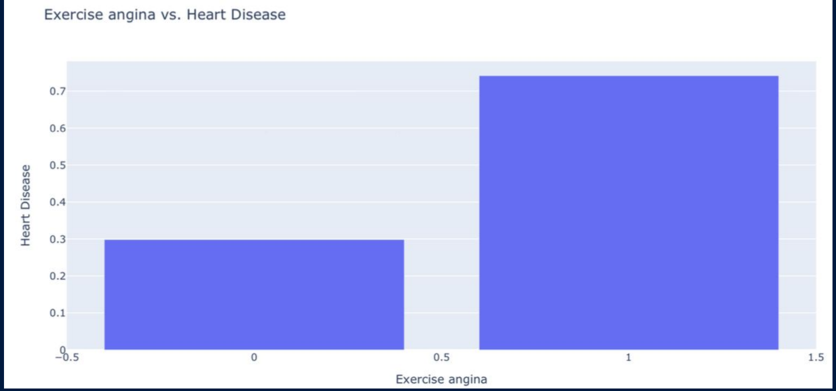
## MAX HR (Maximum Heart Rate)

- **Average Maximum Heart Rate:**
  - Without heart disease (0): ~160 bpm.
  - With heart disease (1): ~140 bpm.
- Individuals with heart disease have a lower average maximum heart rate, suggesting a potential reduction in heart performance.



## EXERCISE ANGINA

- **Prevalence of Heart Disease by Exercise Angina:**
  - No angina during exercise (0): ~30%.
  - Angina during exercise (1): ~70%.
- Exercise-induced angina is a significant indicator of heart disease risk.

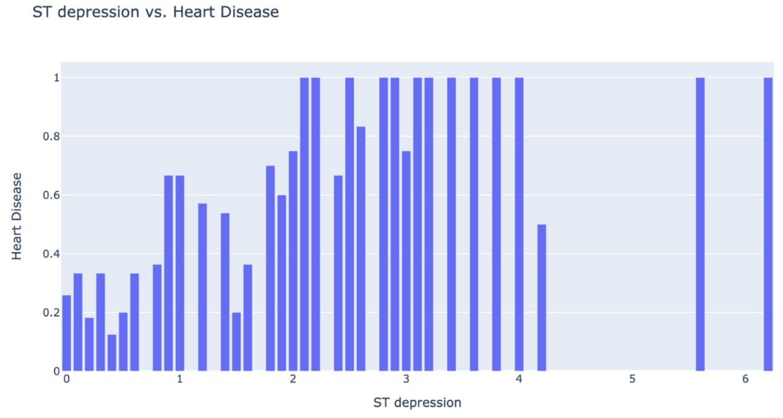


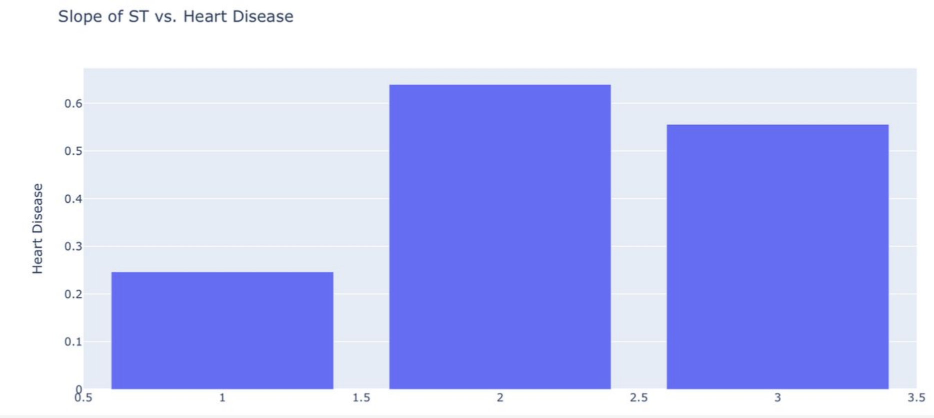
## ST DEPRESSION

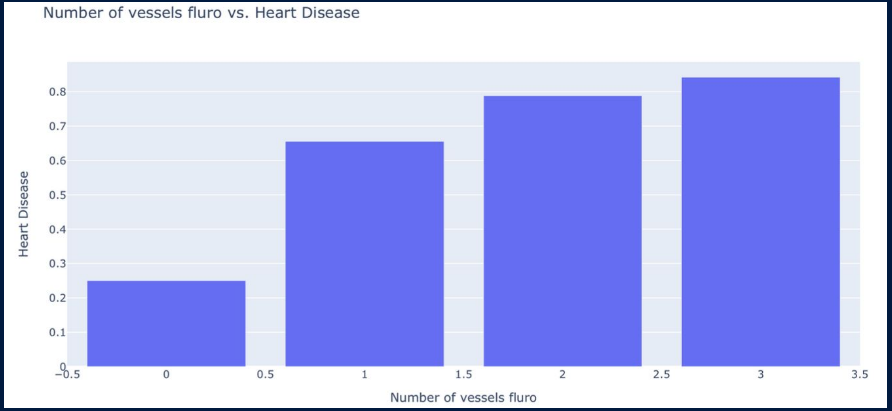
- Heart disease prevalence increases as ST depression values rise.
- ST depression values ≥2 show a high risk of heart disease, ranging from 60%-100%.

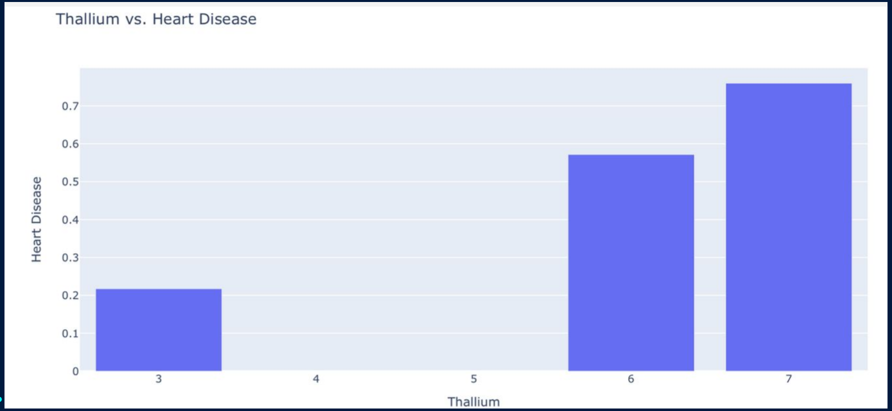
## SLOPE OF ST

- **Relationship Between ST Segment Slope and Heart Disease Risk:**
  - Flat slope: Highest risk.
  - Downward slope: Moderate risk.
  - Upward slope: Lowest risk.









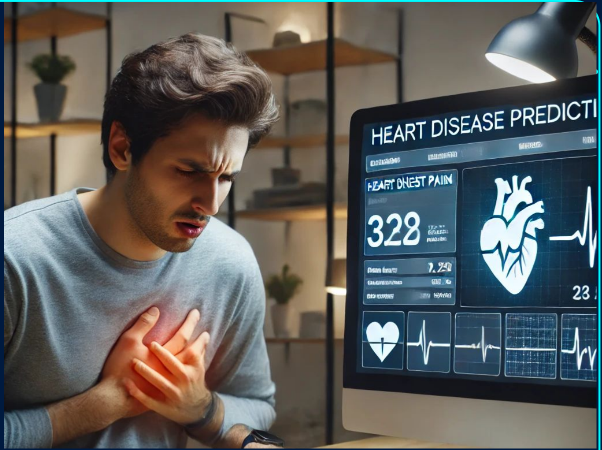
## NUMBER OF VESSELS (Fluoroscopy)

- Heart disease prevalence increases with the number of vessels visible in fluoroscopy:
  - 0 vessels: ~25%.
  - 1 vessel: ~60%.
  - 2 vessels: ~70%.
  - 3 vessels: ~80%.

## THALLIUM (Thallium Test Results)

- **Heart Disease Prevalence by Thallium Test Results:**
  - Result 3: ~20%.
  - Result 6: ~55%.
  - Result 7: ~70%.
- Thallium test results strongly correlate with heart disease risk, with higher results indicating higher risk.

# Results



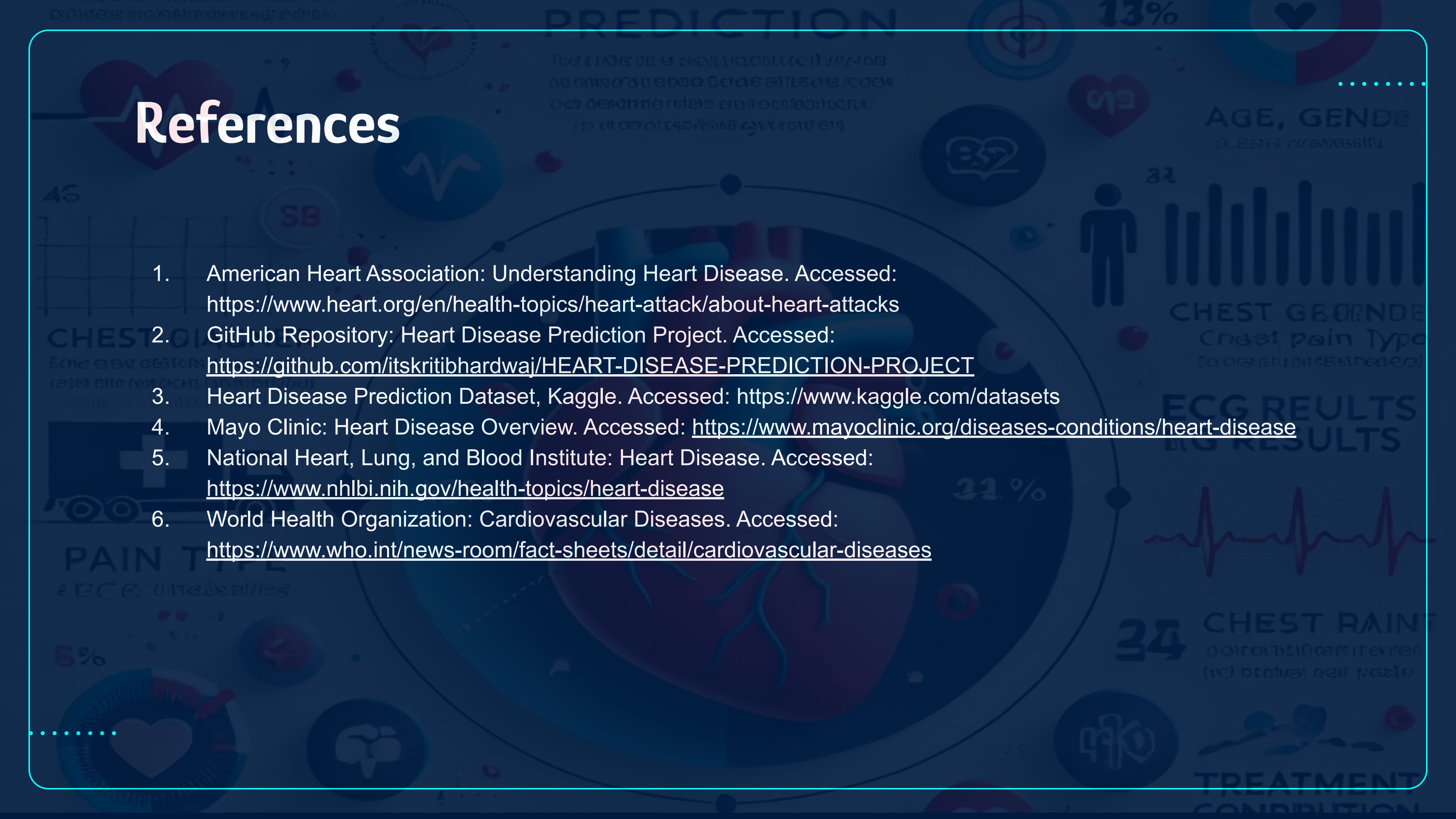
### Analysis and Processes:

- Categorical and numerical variables were identified.
- Analysis of outliers, duplicates, and missing values was performed.
- Data preprocessing was conducted using Label Encoding and Standard Scaler.
- A Random Forest model was used, achieving an accuracy of 85%.
- The model's performance was evaluated using a Confusion Matrix and classification report.

### Findings:

- Features such as age, gender, and chest pain type are significant factors influencing heart disease risk.
- Heart disease risk is associated with age, ECG results, maximum heart rate, exercise-induced angina, ST depression, the number of vessels visible in fluoroscopy, and thallium test results.
- Males have a higher likelihood of developing heart disease compared to females.

**Conclusion:** The model successfully predicts heart disease, and certain demographic and clinical features are crucial in understanding heart disease risk. These findings can contribute to early diagnosis and treatment processes.



# References

1. American Heart Association: Understanding Heart Disease. Accessed: https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks
2. GitHub Repository: Heart Disease Prediction Project. Accessed: https://github.com/itskritibhardwaj/HEART-DISEASE-PREDICTION-PROJECT
3. Heart Disease Prediction Dataset, Kaggle. Accessed: https://www.kaggle.com/datasets
4. Mayo Clinic: Heart Disease Overview. Accessed: https://www.mayoclinic.org/diseases-conditions/heart-disease
5. National Heart, Lung, and Blood Institute: Heart Disease. Accessed: https://www.nhlbi.nih.gov/health-topics/heart-disease
6. World Health Organization: Cardiovascular Diseases. Accessed: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases