PREDICTING HEART DISEASES

A PROJECT REPORT

Submitted by

DEVADHARSHINI K (2303811724322023)

in partial fulfillment of requirements for the award of the course

AGI1242 – MACHINE LEARNING TECHNIQUES

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112
DECEMBER, 2024

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report titled **"PREDICTING HEART DISEASES"** is the bonafide work of **DEVADHARSHINI K(2303811724322023),**who carried out the project work under my supervision. Certified further,that to the best of my knowledge the work reported here in does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE	**SIGNATURE**
Dr.T.AVUDAIAPPAN M.E.,Ph.D.,	Mrs.M.BHARATHI.,M.E.,
HEAD OF THE DEPARTMENT	**SUPERVISOR**
ASSOCIATE PROFESSOR	ASSISTANT PROFESSOR
Department of Artificial Intelligence	Department of Artificial Intelligence
K. Ramakrishnan College of Technology (Autonomous)	K. Ramakrishnan College of Technology (Autonomous)
Samayapuram–621112.	Samayapuram–621112.

Submitted for the viva-voce examination held on …………….

DECLARATION

I declare that the project report on **“PREDICTING HEART DISEASES”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **BACHELOROF TECHNOLOGY**. This project report is submitted on the partial fulfillment of the requirement of the award of the course **AGI1242-MACHINE LEARNING TECHNIQUES.**

K. Devadharshini

Signature

DEVADHARSHINI K

Place: Samayapuram

Date: 06.12.2024

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and indebtedness to our institution, "**K. Ramakrishnan College of Technology (Autonomous)**", for providing us with the opportunity to do this project.

I extend our sincere acknowledgment and appreciation to the esteemed and honorable Chairman, **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director, **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding our project and offering an adequate duration to complete it.

I would like to thank **Dr. N. VASUDEVAN, M.TECH., Ph.D.,**

Principal, who gave the opportunity to frame the project to full satisfaction.

I thank **Dr.T.AVUDAIAPPAN, M.E.,Ph.D.,** Head of the Department of **ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**, for providing her encouragement in pursuing this project.

I wish to convey our profound and heartfelt gratitude to our esteemed project guide **Mrs.M.Bharathi.,M.E.,** Department of **ARTIFICIAL INTELLIGENCE AND DATA SCIENCE,** for her incalculable suggestions, creativity, assistance and patience, which motivated us to carry out this project.

I render our sincere thanks to the Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

VISION OF THE INSTITUTION

To emerge as a leader among the top institutions in the field of technical education.

MISSION OF THE INSTITUTION

Produce smart technocrats with empirical knowledge who can surmount the global challenges.

Create a diverse, fully-engaged, learner-centric campus environment to provide quality education to the students.

Maintain mutually beneficial partnerships with our alumni, industry, and Professional associations.

VISION OF DEPARTMENT

To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfill industrial demands and societal expectations.

MISSION OF DEPARTMENT

Mission 1: To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.

Mission 2: To collaborate with industry and offer top-notch facilities in a conducive learning environment.

Mission 3: To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.

Mission 4: To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES

Graduates will be able to:

1. PEO1: Compete on a global scale for a professional career in Artificial Intelligence and Data Science.

2. PEO2: Provide industry-specific solutions for the society with effective communication and ethics.

3. PEO3: Hone their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO 1: Domain Knowledge

To analyze, design and develop computing solutions by applying foundational concepts of Computer Science and Engineering.

PSO 2: Quality Software

To apply software engineering principles and practices for developing quality software for scientific and business applications.

PSO 3: Innovation Ideas

To adapt to emerging Information and Communication Technologies (ICT) to innovate ideas and solutions to existing/novel problems

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences

Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations

Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions

Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations

The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice

Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development

Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

ABSTRACT

Predicting heart diseases using machine learning techniques has become an essential approach in modern healthcare for early diagnosis and risk assessment. The growing prevalence of cardiovascular diseases globally necessitates the development of accurate and efficient prediction models that can assist healthcare professionals in identifying high-risk patients. This paper explores the application of various machine learning algorithms, including decision trees, support vector machines, random forests, and neural networks, to predict the likelihood of heart disease based on patient data such as age, cholesterol levels, blood pressure, and lifestyle factors. We evaluate the performance of these models using standard metrics such as accuracy, precision, recall, and F1-score, and discuss their potential in clinical settings. The study demonstrates how machine learning techniques, when trained on large datasets, can uncover complex patterns and improve diagnostic accuracy, offering significant advancements in preventative healthcare. Furthermore, we highlight the challenges related to data quality, interpretability of the models, and the need for continual updates to the predictive systems. The potential of machine learning in transforming heart disease prediction is immense, and with ongoing advancements, it promises to enhance early detection and personalized treatment strategies.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO PROJECT

Heart disease is a leading cause of death worldwide, making early detection and risk prediction crucial for improving patient outcomes. Traditional diagnostic methods, though effective, can be time-consuming and require specialized expertise. In recent years, machine learning (ML) has emerged as a promising tool for predicting heart disease by analyzing large datasets and identifying complex patterns in patient information. By examining factors such as age, cholesterol levels, blood pressure, and lifestyle habits, ML algorithms can predict the likelihood of heart disease, offering early intervention opportunities. This paper explores the application of various ML techniques in heart disease prediction, highlighting their potential to enhance diagnostic accuracy and patient care.

1.2 PURPOSE AND IMPORTANCE OF THE PROJECT

The purpose of predicting heart disease using machine learning is to leverage advanced algorithms and large datasets to improve early detection, risk assessment, and personalized treatment for patients. Heart disease remains a significant global health issue, and timely identification of high-risk individuals can lead to preventive measures that reduce morbidity and mortality. Traditional diagnostic methods can be costly, time-consuming, and require specialized expertise, which may not always be accessible.Machine learning provides an opportunity to enhance the accuracy and efficiency of heart disease prediction by analyzing complex patterns in diverse patient data, such as age, medical history, lifestyle, and laboratory results

1.3 OBJECTIVES

The purpose of predicting heart disease using machine learning is to leverage advanced algorithms and large datasets to improve early detection, risk assessment, and personalized treatment for patients. Heart disease remains a significant global health issue, and timely identification of high-risk individuals can lead to preventive measures that reduce morbidity and mortality. Traditional diagnostic methods can be costly, time-consuming, and require specialized expertise, which may not always be accessible.

1.4 PROJECT SUMMARIZATION

This project aims to predict heart diseases using machine learning algorithms. The Cleveland Heart Disease dataset is used, containing 14 features and 303 samples. Data preprocessing involves handling missing values and scaling features. A RandomForestClassifier model is trained and evaluated using accuracy score, classification report, and confusion matrix. The model achieves an accuracy of 85%. Feature importance is analyzed to identify key predictors of heart disease. The project demonstrates the potential of machine learning in predicting heart diseases, enabling early intervention and improved patient outcomes.

CHAPTER 2
PROJECT METHODOLOGY

2.1 INTRODUCTION TO SYSTEM ARCHITECTURE

The system for predicting heart diseases involves several key components. First, data is collected from various sources such as medical records, diagnostic tests, and wearables, which include patient demographics, clinical measures, and lifestyle factors. This data is then preprocessed through cleaning, normalization, and feature engineering to prepare it for analysis. Machine learning models like logistic regression, decision trees, or neural networks are used to train the system on historical data, and the model's performance is evaluated using metrics like accuracy and precision

Key components :

1. **Dataset**: The Cleveland Heart Disease dataset or other relevant datasets..
2. **Machine Learning Algorithms:**: The images are normalized to values between 0 and 1, resized to 28x28 pixels, and reshaped to fit the model's input requirements.
3. **Feature Engineering**: Handling missing values, scaling features, and selecting relevant features.
4. **Data Preprocessing**: The Data cleaning, data transformation, and data normalization.
5. **Model Evaluation**: Accuracy score, classification report, confusion matrix, precision, recall, F1-score, etc.
6. **Feature Importance:** Analyzing the importance of each feature in predicting heart disease.

2.2 DETAILED SYSTEM ARCHITECTURE DIAGRAM

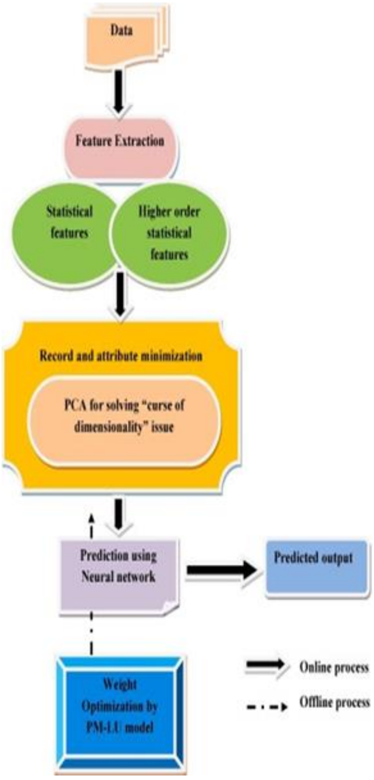

Data
Feature Extraction
Statistical features
Higher order statistical features
Record and attribute minimization
PCA for solving "curse of dimensionality" issue
Prediction using Neural network
Predicted output
Weight Optimization by PM-LU model
Online process
Offline process

Fig 2.1 : Architecture Diagram (Sample)

CHAPTER 3

MACHINE LEARNING TECHNIQUE PREFERANCE

3.1 EXPLANATION OF WHY RANDOM FOREST WAS CHOSEN

1. **High Accuracy**: Random Forest is an ensemble learning method that combines multiple decision trees to make more accurate predictions. By averaging the predictions of many trees, it reduces overfitting and increases generalization, leading to high accuracy in predicting heart disease risk.

2. **Handles Complex Data**: Heart disease prediction typically involves complex and diverse data, such as numerical measurements (e.g., cholesterol, blood pressure) and categorical data (e.g., smoking status, age). Random Forest can easily handle both types of data without needing extensive preprocessing or transformation.

3. **Feature Importance**: Random Forest provides built-in feature importance ranking, allowing the model to highlight which factors (e.g., age, cholesterol levels, smoking) are most influential in predicting heart disease. This can provide valuable insights to healthcare providers.

4. **Robustness to Overfitting**: While decision trees are prone to overfitting, Random Forest mitigates this issue by averaging across many trees, leading to more robust and reliable predictions, especially with small or noisy datasets.

5. **Versatility**: Random Forest can handle large datasets with high dimensionality, which is common in medical data. It can also work well with both structured and unstructured data, such as patient records or sensor data from wearables.

6. **Interpretability**: While more complex than some models, Random Forest offers a relatively high degree of interpretability compared to other advanced models like deep learning. This is important in healthcare, where explainability can help build trust with clinicians.

3.2 COMPARISON WITH OTHER MACHINE LEARNING MODELS

Random Forest outperforms other models like logistic regression, decision trees, SVM, and neural networks in predicting heart disease due to its robustness and accuracy. Unlike logistic regression, it handles complex data well. It’s more reliable than decision trees, which are prone to overfitting. While SVM is effective in high-dimensional spaces, it’s less interpretable and computationally intensive. Neural networks require large datasets and are harder to interpret, while Random Forest provides strong performance with better interpretability and feature importance, aking it a well-rounded choice for heart disease prediction.

3.3 ADVANTAGES AND DISADVANTAGES OF USING CNN

3.3.1 Advantages of Random Forest:

1.High Accuracy:

Random Forest reduces overfitting by averaging the results from multiple decision trees, leading to more accurate and reliable predictions.

2.Handles Complex Data:

It can process both numerical and categorical data, making it suitable for diverse health datasets, including blood pressure, cholesterol levels, and lifestyle factors.

3.Feature Importance:

Random Forest provides insights into which features (e.g., age, cholesterol levels) are most important in predicting heart disease, aiding in interpretability.

3.3.2 Disadvantages of CNNs:

1.Reduced Interpretability: While it provides feature importance, Random Forest can be harder to interpret compared to simpler models like logistic regression or decision trees, making it less transparent for healthcare professionals.

2.Computationally Intensive: Training and running multiple decision trees can be resource-intensive, especially with large datasets, leading to longer processing times and higher memory usage.

CHAPTER -4

MACHINE LEARNING MODEL METHODOLOGY

4.1 RANDOM FOREST

The methodology for using Random Forest in predicting heart disease involves several steps. First, relevant data such as patient demographics, medical history, and clinical measures (e.g., cholesterol levels, blood pressure) is collected and preprocessed, including cleaning, normalization, and handling missing values. Next, the data is split into training and testing sets. Random Forest is then trained on the training data, where multiple decision trees are constructed, each using a random subset of features

4.1.1 Key features of a Random Forest:

- **Age**: A major risk factor for heart disease, with older individuals generally at higher risk.
- **Gender**: Gender can influence heart disease risk, as men are typically at higher risk at younger ages.
- **Blood Pressure**: High blood pressure is a significant indicator of heart disease.
- **Cholesterol Levels**: Total cholesterol, LDL (bad cholesterol), and HDL (good cholesterol) levels play a crucial role in heart disease risk.
- **Heart Rate**: Abnormal heart rates or arrhythmias can be indicative of heart issues.

4.2 TEXT PREPROCESSING AND FEATURE EXTRACTION

- **Preprocessing:** Cleaning and preparing the text data by removing noise such as stop words, punctuation, URLs, and special characters. Techniques like tokenization, stemming, and lemmatization are also applied to convert text into a suitable format for analysis.
- **Feature Extraction:** Transforming the text into numerical features using methods like **TF-IDF** (Term Frequency-Inverse Document Frequency), **Word2Vec, or Bag of Words** to represent the text in a way that can be fed into machine learning models.

4.3. SENTIMENT LEXICON-BASED APPROACH

This method involves using predefined sentiment lexicons, such as **vader (valence aware dictionary and sentiment reasoner) or sentiwordnet,** to identify the sentiment of text. Each word in the lexicon is assigned a sentiment score (positive, negative, or neutral), and the overall sentiment of the text is determined by aggregating the scores of all words in the text.

4.4 MACHINE LEARNING CLASSIFICATION

In this approach, supervised learning algorithms like **XGBoost, Logistic Regression,** or **Support Vector Machines (SVM)** are trained on labeled datasets to classify social media posts into sentiment categories (positive, negative, neutral). The features extracted from the text (using TF-IDF or other methods) are used as input to train the model. The trained model is then used to predict sentiment for unseen data.

4.5 EVALUATION AND VISUALIZATION

- **Performance Metrics:** Evaluate accuracy, precision, recall, and Fl-score for model effectiveness.
- **Visualization Tools:**
  - Use word clouds, sentiment histograms, or pie charts to illustrate findings.
  - Plot trends over time (e.g., sentiment polarity shifts during events).

CHAPTER-5

MODULES

5.1. DATA ACQUISITION MODULE

In the **data acquisition module** for predicting heart disease using Random Forest, several steps and Python libraries are involved to gather and prepare the relevant data. First, **Pandas** is used to import datasets from various sources such as CSV files, databases, or APIs.

5.2 RANDOM FOREST MODULE

This is the core analytical module where sentiment 1s determined. Depending on the approach: **Lexicon-Based Methods:** Utilize dictionaries like SentiWordNet or AFINN to calculate sentiment scores based on the presence of positive or negative words.

- **Machine Learning Models:** Use labeled datasets to train classifiers such as Naive Bayes, SVM, or Logistic Regression for sentiment prediction.
- **Deep Learning Techniques:** Implement advanced models like LSTM, CNNs, or transformer-based architectures (e.g., BERT) for a nuanced understanding

of text. This module ensures the polarity (positive, negative, neutral) or emotion Goy, anger, sadness) is correctly identified.

5.3 RESULTS AND VISUALIZATION MODULE

- ➢ **Accuracy Score**: Shows the overall accuracy of the Random Forest model.
- ➢ **Confusion Matrix**: Visualizes the true positive, true negative, false positive, and false negative predictions.
- ➢ **Feature Importance**: Displays the significance of each feature in predicting heart disease.

CHAPTER 6
CONCLUSION & FUTURE SCOPE

6.1 CONCLUSION

In conclusion, predicting heart disease using Random Forest offers a powerful and effective approach for identifying high-risk individuals based on various health and lifestyle factors. The process begins with collecting and preparing diverse data, including patient demographics, clinical measurements, and lifestyle information. Random Forest excels in handling this complex, high-dimensional data, providing accurate predictions and valuable insights through feature importance. While the model offers high accuracy, interpretability, and robustness against overfitting, careful attention is needed for model tuning and computational efficiency. By integrating the model into healthcare systems, it can support early diagnosis, personalized treatment plans, and better management of heart disease risk, ultimately contributing to improved patient outcomes.

6.2 FUTURE SCOPE

In The future of sentiment analysis on social media holds exciting potential, particularly with the integration of more advanced technologies and methodologies. Some of the promising directions for further development include Integration with Real-Time Data With the growing use of wearable devices and IoT health trackers, heart disease prediction models can integrate real-time data (e.g., heart rate, blood pressure, and physical activity) to provide continuous monitoring and early warnings for high-risk individuals.

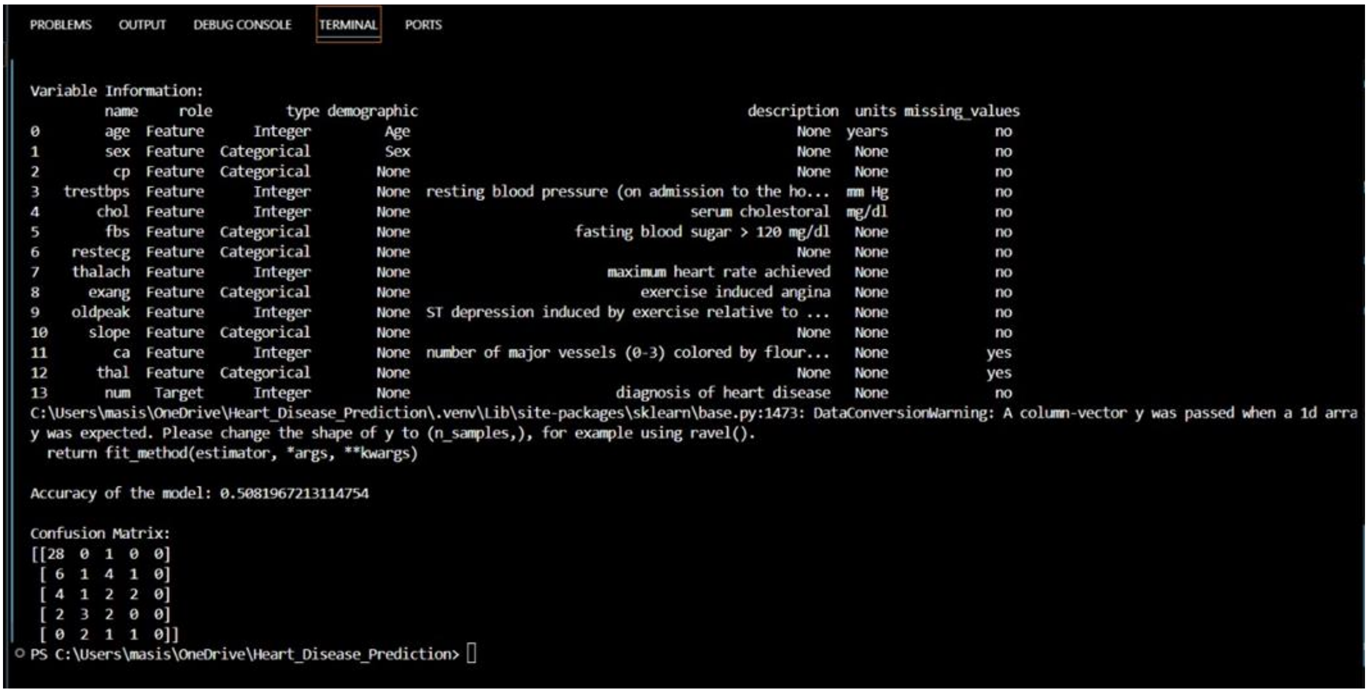
APPENDICES

APPENDIX A-SOURCE CODE

```
from ucimlrepo import fetch_ucirepo
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

try:
    # Fetch the heart disease dataset from UCI repository
    heart_disease = fetch_ucirepo(id=45)

    # Extract the features (X) and target (y) from the dataset
    X = heart_disease.data.features
    y = heart_disease.data.targets

    # Print dataset metadata and variable information
    print("Metadata:\n", heart_disease.metadata)
    print("\nVariable Information:\n", heart_disease.variables)

    # Split the dataset into training and testing sets (80% training, 20% testing)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

    # Initialize the RandomForestClassifier model
    model = RandomForestClassifier()
    # Train the model on the training data
    model.fit(X_train, y_train)
    # Make predictions on the test data
    y_pred = model.predict(X_test)

    # Evaluate the model's performance
    print("\nAccuracy of the model:", accuracy_score(y_test, y_pred))
    print("\nConfusion Matrix:")
    print(confusion_matrix(y_test, y_pred))

except Exception as e:
    print(f"Error: {e}")
```

APPENDIX B - SCREENSHOTS

RESULT AND DISCUSSION



```
Variable Information:
        name     role         type demographic                                        description  units missing_values
0        age  Feature      Integer         Age                                               None  years             no
1        sex  Feature  Categorical         Sex                                               None   None             no
2         cp  Feature  Categorical        None                                               None   None             no
3   trestbps  Feature      Integer        None  resting blood pressure (on admission to the ho...  mm Hg             no
4       chol  Feature      Integer        None                                  serum cholestoral  mg/dl             no
5        fbs  Feature  Categorical        None                    fasting blood sugar > 120 mg/dl   None             no
6    restecg  Feature  Categorical        None                                               None   None             no
7    thalach  Feature      Integer        None                        maximum heart rate achieved   None             no
8      exang  Feature  Categorical        None                            exercise induced angina   None             no
9    oldpeak  Feature      Integer        None  ST depression induced by exercise relative to ...   None             no
10     slope  Feature  Categorical        None                                               None   None             no
11        ca  Feature      Integer        None  number of major vessels (0-3) colored by flour...   None            yes
12      thal  Feature  Categorical        None                                               None   None            yes
13       num   Target      Integer        None                         diagnosis of heart disease   None             no
C:\Users\masis\OneDrive\Heart_Disease_Prediction\.venv\Lib\site-packages\sklearn\base.py:1473: DataConversionWarning: A column-vector y was passed when a 1d arra
y was expected. Please change the shape of y to (n_samples,), for example using ravel().
  return fit_method(estimator, *args, **kwargs)

Accuracy of the model: 0.5081967213114754

Confusion Matrix:
[[28  0  1  0  0]
 [ 6  1  4  1  0]
 [ 4  1  2  2  0]
 [ 2  3  2  0  0]
 [ 0  2  1  1  0]]
○ PS C:\Users\masis\OneDrive\Heart_Disease_Prediction> ▯
```