

Heart Disease Prediction Using Machine Learning Classification Models

Project By: Yash Mhaske

Important Links:

Colab:

<https://colab.research.google.com/drive/1aTEckmGbzUjd3jQd2yLiA9MajrjmWUud?usp=sharing>

Github:

<https://github.com/yash11103/HEART-DISEASE-PREDICTION-MAJOR-PROJECT>

Gradio:

<https://16280d13e604f00f65.gradio.live>

1.Introduction

This project focuses on predicting whether a person has heart disease or not. Multiple models are used for this purpose and compared. This will interpret feature importance and understand key factors contributing to heart disease.

2.Problem Statement

Heart disease remains one of the leading causes of death worldwide. Early and accurate prediction can significantly improve patient outcomes and reduce mortality. Traditional diagnosis methods are time-consuming and often require expensive tests.

The goal is to build and compare several classification models such as Logistic Regression, Decision Tree, Random Forest, and AI-based methods (e.g., Neural Networks) to predict the presence of heart disease based on the given features.

3.Objective

To develop a supervised classification model that predicts whether a patient has heart disease (target: 0 or 1).

To compare the performance of multiple models:

Logistic Regression, Decision Tree, Random Forest, AI model (e.g., simple Feedforward Neural Network)

To evaluate models using classification metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

4.Dataset Description:

Dataset:<https://drive.google.com/file/d/1k3Yhgzrgzl9CbdGXuZvK7WgbZ8kVx56I/view?usp=sharing>

usp=sharing

The dataset contains numerical data regarding various features about studying the conditions for heart disease like thal, cp, trestbps, etc.

5.Methodology:

We used Logistic Regression, Decision Tree, Random forest and Neural Networks which are supervised machine learning models to predict the heart disease probability. The dataset was split into training(80%) and testing(20%) sets. The models were evaluated using accuracy, precision, recall, F1 score.

6.Model Evaluation:

1.Logistic Regression: Test Accuracy: 0.8868 (88.68%)

2.Decision Tree: Test Accuracy: 0.8491 (84.91%)

3.Random Forest: Test Accuracy: 0.8302 (83.02%)

4.Neural Network: Test Accuracy: 0.6066 (60.66%)

Logistic Regression model is the best performing model among these four.

Confusion matrix: $\begin{bmatrix} 19 & 4 \\ 2 & 28 \end{bmatrix}$

Classification Report

0: precision=0.90, recall=0.83, f1-score=0.86, support=23

1: precision=0.88, recall=0.93, f1-score=0.90, support=30

Macro avg: precision=0.89, recall=0.88,

f1-score=0.88 support=53
weighted avg: precision=0.89, recall=0.89,
f1score=0.89 support=53

7.Results and Insights

The logistic regression model achieved a high accuracy, demonstrating strong ability to predict the presence of heart disease. Higher 'oldpeak' values are more prevalent in patients with heart disease. Lower 'oldpeak' values are more common in patients without heart disease.

8.Target Audience

Hospitals and clinics for pre-screening patients

Health insurance providers for risk assessment

Data scientists and healthcare professionals interested in preventive analytics

9.Conclusion

We have implemented the supervised learning models to predict the presence of heart diseases, evaluated those models and deployed it using gradio.