Project Title: Heart Disease Prediction using Machine Learning

Name: Raniya Shareef
Course: AI with Python
Institution: Techmaghi
Date: 29 June 2025

1. Introduction

In this project, I explored how machine learning can be used to predict whether a person has heart disease based on their medical data. Early detection can really help in preventing serious health issues. I used a public dataset and some basic ML models to see how accurately we can make predictions.

2. Data Preprocessing

> **Note:** Due to limitations of the online compiler used (OnlineGDB), only a small sample of the dataset was used in this project. This is sufficient to demonstrate model training and evaluation, but not intended for real-world deployment.

- **Dataset Overview:**
- Total Records: 1,025
- Features: 14 (both categorical and numerical)
- Target: Heart disease (1 = present, 0 = not present)
- **Missing Values:**
- Luckily, the dataset didn't have missing values.
- **Outlier Handling:**
- I checked for outliers using box plots, especially in columns like cholesterol and 'oldpeak'.
- **Encoding Categorical Data:**
- Categorical columns like 'cp', 'thal', and 'slope' were turned into numbers using label encoding.
- **Feature Scaling:**
- I used StandardScaler to scale features like age, cholesterol, and resting blood pressure. This helps improve model accuracy.

3. Exploratory Data Analysis (EDA)

- I made **histograms** to check distributions of variables like age and cholesterol.
- **Box plots** helped me find outliers.

- A **correlation heatmap** showed which features are related to each other. For example, chest pain type and maximum heart rate had a strong relation to the target.
- The target variable was pretty balanced: 526 patients had heart disease, and 499 didn't.

4. Model Selection & Training

I tried three different machine learning models: - **Logistic Regression** – simple and easy to understand. - **Support Vector Machine (SVM)** – useful for classification tasks. - **Random Forest** – combines many decision trees and usually gives good accuracy.

I split the data into training and test sets (80/20) and trained each model separately. I also used cross-validation to make the results more reliable.

5. Evaluation & Results

Here's how the models performed:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83%	82%	84%	83%
SVM	85%	84%	85%	84%
Random Forest	**88%**	87%	89%	**88%**

Random Forest was the best overall, with the highest scores.

I also used a **confusion matrix** and **ROC curve** to understand the performance visually.

6. Conclusion

This project helped me understand how machine learning can be used in health care. The Random Forest model gave the best results. In the future, I would like to try more advanced techniques like tuning the model or even deploying it online. While it's not perfect, it's a good starting point for helping with early detection of heart disease.

7. Project Evidence

- Screenshot of code and output is attached.
- Google Drive link to project files: https://drive.google.com/drive/folders/137ry3Y4SKbcDU7YhzS3Ttd4DyY6Cble7?usp=sharing

End of Report

https://drive.google.com/drive/folders/137ry3Y4SKbcDU7YhzS3Ttd4DyY6Cble7?usp=drive_link