Heart Disease Prediction using Machine Learning

Internship Project: Digital Empowerment Network (Week 1)

Mentor: Hussain Shoaib

Student: Zarar Shahid

📊 Project Objective

To understand and implement a full machine learning pipeline on a real-world classification problem: predicting heart disease presence using clinical data.

☑️ Step-by-Step Workflow

☑️ 1. Dataset Used

- Name: Heart Disease UCI Dataset
- Source: Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci)
- Shape: 303 rows × 14 columns

☑️ 2. Data Preprocessing

- Missing Values: None
- Feature Scaling: StandardScaler used to normalize all features
- Target Variable: `target` (1 = disease, 0 = healthy)

☑️ 3. Exploratory Data Analysis (EDA)

- Correlation heatmap revealed strong relationships:
  - `cp` (chest pain type), `thalach` (max heart rate), and `exang` (exercise-induced angina) showed strong correlation with target

☑ 4. Model Training & Evaluation

Split: 80% train / 20% test

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	~0.77	~0.78	~0.82	~0.80
Random Forest	~0.85	~0.85	~0.89	~0.87
Logistic Regression	~0.86	~0.87	~0.87	~0.87
SVM	~0.87	~0.89	~0.87	~0.88

☑ 5. Hyperparameter Tuning (Bonus)

- Model: Random Forest
- Method: GridSearchCV
- Best Params: `n_estimators=100`, `max_depth=5`
- Result: Accuracy improved to ~0.88

☑ 6. Model Export

- Best model saved using `joblib` as `best_rf_model.pkl`

🏆 Outcome & Learnings

- Understood full ML cycle: preprocessing → EDA → modeling → evaluation
- Gained hands-on with multiple algorithms (Decision Tree, Random Forest, Logistic Regression, SVM)
- Practiced GridSearchCV and model saving for deployment

✉ Future Work

- Add ROC-AUC and cross-validation
- Deploy model using Streamlit web app
- Improve interpretability with SHAP or LIME

📅 Submission Date

July 23, 2025

Prepared by: Zarar Shahid
Mentor: Hussain Shoaib