Disease Prediction Using Patient Data

Objective

To train and evaluate machine learning models to predict heart disease based on patient data.

Dataset Overview

- **Source**: `heart.csv` dataset
- **Features**:
  - Numerical: `age`, `trestbps`, `chol`, `thalach`, `oldpeak`
  - Categorical: `cp`, `restecg`, `slope`, `thal`, `sex`, `fbs`, `exang`
- **Target**: `target` (1: Disease, 0: No Disease)
- **Dataset Size**:
  - Total Samples: 1025
  - Training Set: 820 samples
  - Test Set: 205 samples

Methodology

1. Data Preprocessing

- **Handling Missing Values**: None present in this dataset.
- **Scaling**: Standardized numerical features.
- **Encoding**: One-hot encoded categorical features, resulting in 25 features post-encoding.

2. Model Training

Three machine learning models were trained:

- **Logistic Regression**: Baseline linear model.
- **Random Forest**: Ensemble learning method for high accuracy.
- **Support Vector Machine (SVM)**: Non-linear model for better classification.

3. Evaluation Metrics

- **Precision**: Ratio of true positives to predicted positives.
- **Recall**: Ratio of true positives to actual positives.
- **F1-Score**: Harmonic mean of precision and recall.
- **Support**: Total number of samples.

Results

Model	Precision	Recall	F1-Score
Logistic Regression	77.1%	77.1%	77.1%
Random Forest	100.0%	100.0%	100.0%
SVM	87.98%	87.80%	87.79%

Insights

- **Random Forest** achieved perfect performance, indicating possible overfitting.
- **SVM** offers a balanced trade-off between complexity and performance.
- **Logistic Regression** provides a simple but less accurate baseline.

Recommendations

- Use **Random Forest** for scenarios prioritizing high accuracy.
- Employ **SVM** for a robust and generalizable model.
- Further validation with cross-validation is recommended to confirm Random Forest's performance.

Deliverables

1. **Code Repository**:
    - Preprocessing pipeline.
    - Training and evaluation scripts for Logistic Regression, Random Forest, and SVM.
2. **Presentation Slides**:
    - Summary of dataset, analysis, and model performance.
3. **Insights Document**:
    - Detailed analysis and recommendations for future improvements.

Next Steps

- Conduct cross-validation.
- Hyperparameter optimization for SVM and Random Forest.
- Deployment pipeline for real-world predictions.