Heart Disease Prediction Using Random Forest

Domain Selection:
Healthcare / Medical Diagnosis. The system focuses on predicting heart disease based on patient clinical parameters, helping doctors identify at-risk patients early.

Project Objective:
Build a system that predicts whether a patient has heart disease using Random Forest, leveraging structured patient data such as vitals, lab results, and ECG readings.

Dataset:
Contains patient information including age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, exercise induced angina, ST depression, slope, number of vessels colored, and thalassemia. The target variable indicates presence (1) or absence (0) of heart disease.

Data Preprocessing:
- Outliers handled using IQR method
- No scaling required for Random Forest
- No PCA, LDA, or RFE applied to preserve interpretability
- All numerical and categorical variables used appropriately for modeling

Model Selection:
- Random Forest Classifier
- Hyperparameter tuning using GridSearchCV with 5-fold cross-validation
- Optimized for Recall due to medical importance (minimizing false negatives)

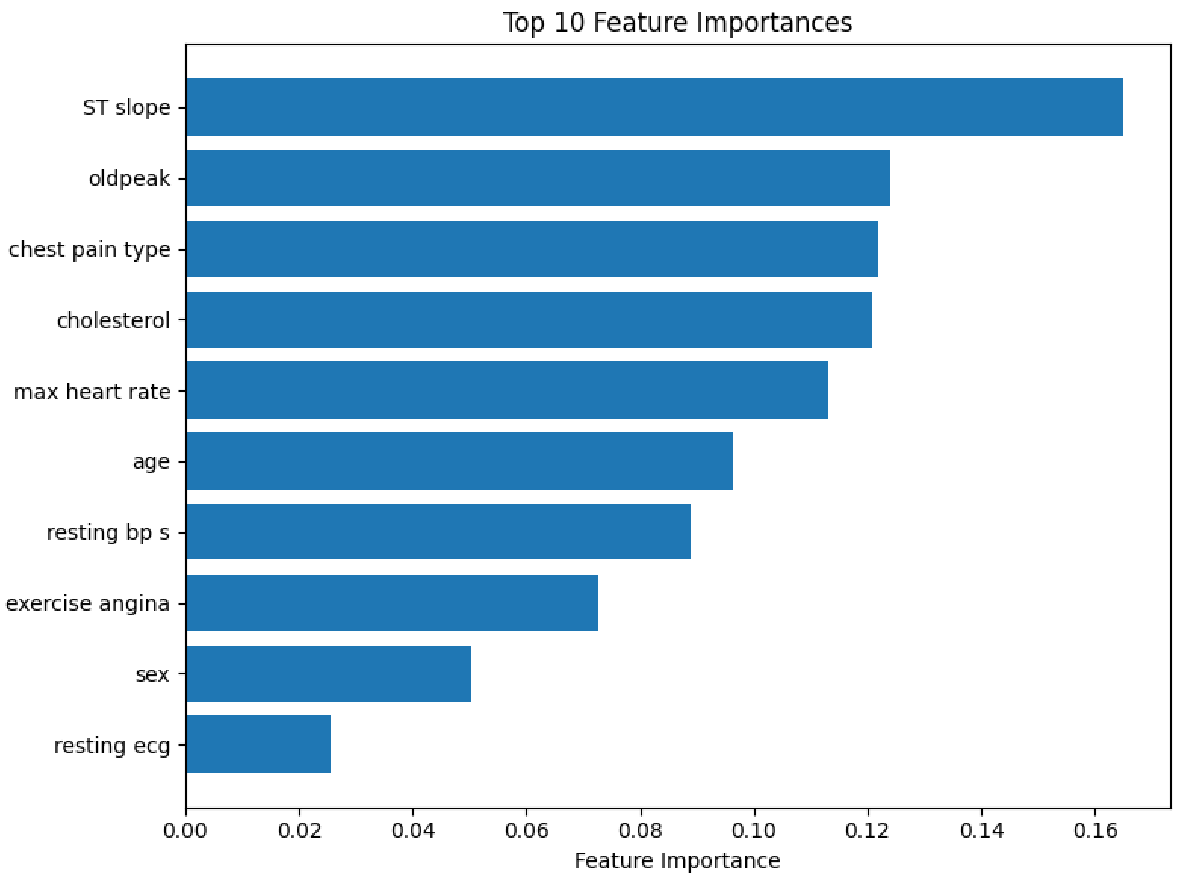
Best Hyperparameters:
criterion: entropy
max_depth: None
max_features: sqrt
min_samples_split: 5
min_samples_leaf: 1
n_estimators: 100

Model Performance on Test Data:
Accuracy: 0.9202

Classification Report:
Precision, Recall, F1-Score for each class:
Class 0 (No Heart Disease) -> precision: 0.91, recall: 0.92, f1: 0.92
Class 1 (Heart Disease) -> precision: 0.93, recall: 0.92, f1: 0.92

Confusion Matrix:
[[103, 9], [10, 116]]

Feature Importance: See figure below



Sample Prediction:
Input Patient Data Example:
{'age': 55, 'sex': 1, 'cp': 2, 'trestbps': 140, 'chol': 250, 'fbs': 0, 'restecg': 1, 'thalach': 150, 'exang': 0, 'oldpeak': 1.5, 'slope': 1, 'ca': 0, 'thal': 2}
Prediction: No Heart Disease
Probability: [[0.77895, 0.22105]]