DECISION TREE & RANDOM FOREST REPORT

In this project, we explored the use of tree-based machine learning models—**Decision Tree Classifier** and **Random Forest Classifier**—to predict the likelihood of heart disease in patients using a structured dataset. The objective was to develop interpretable and accurate classification models, analyze their performance, and understand which features contribute most to heart disease prediction.

📊 Dataset

The dataset, contains medical attributes such as:
- Age, sex, chest pain type (cp)
- Resting blood pressure (trestbps)
- Serum cholesterol (chol)
- Fasting blood sugar, ECG results, max heart rate (thalach)
- ST depression (oldpeak), and others.

The target variable is binary:
0: No heart disease
1: Presence of heart disease

🌳 Decision Tree Classifier

We trained a **Decision Tree** with limited depth (max_depth=3) to avoid overfitting. The tree provides a clear visualization of how features like chest pain type (cp) and ST depression (oldpeak) guide predictions.

We analyzed overfitting by plotting accuracy across different tree depths. Deeper trees performed better on training data but worse on test data, confirming overfitting beyond depth ~4.

🌲 Random Forest Classifier

Next, we trained a **Random Forest** with 100 trees and controlled depth. It performed better than a single decision tree in both accuracy and stability:

Model	Test Accuracy	Cross − Validation Accuracy
Decision Tree	~80%	~79%
Random Forest	~87%	~83%

🍁 Feature Importance

Using the Random Forest model, we extracted and visualized feature importance:
- **Chest Pain Type (cp)** – Most significant; certain chest pain types strongly correlate with heart disease.
- **Max Heart Rate (thalach)** – Lower heart rate after exercise can indicate poor cardiac function.
- **ST Depression (oldpeak)** – Indicates potential myocardial ischemia.
- **Number of Major Vessels (ca)** – Greater number of affected vessels increases disease likelihood.

Conclusion:

Dataset : "C:\Users\User\Downloads\archive (7)\heart_disease_dataset.csv"
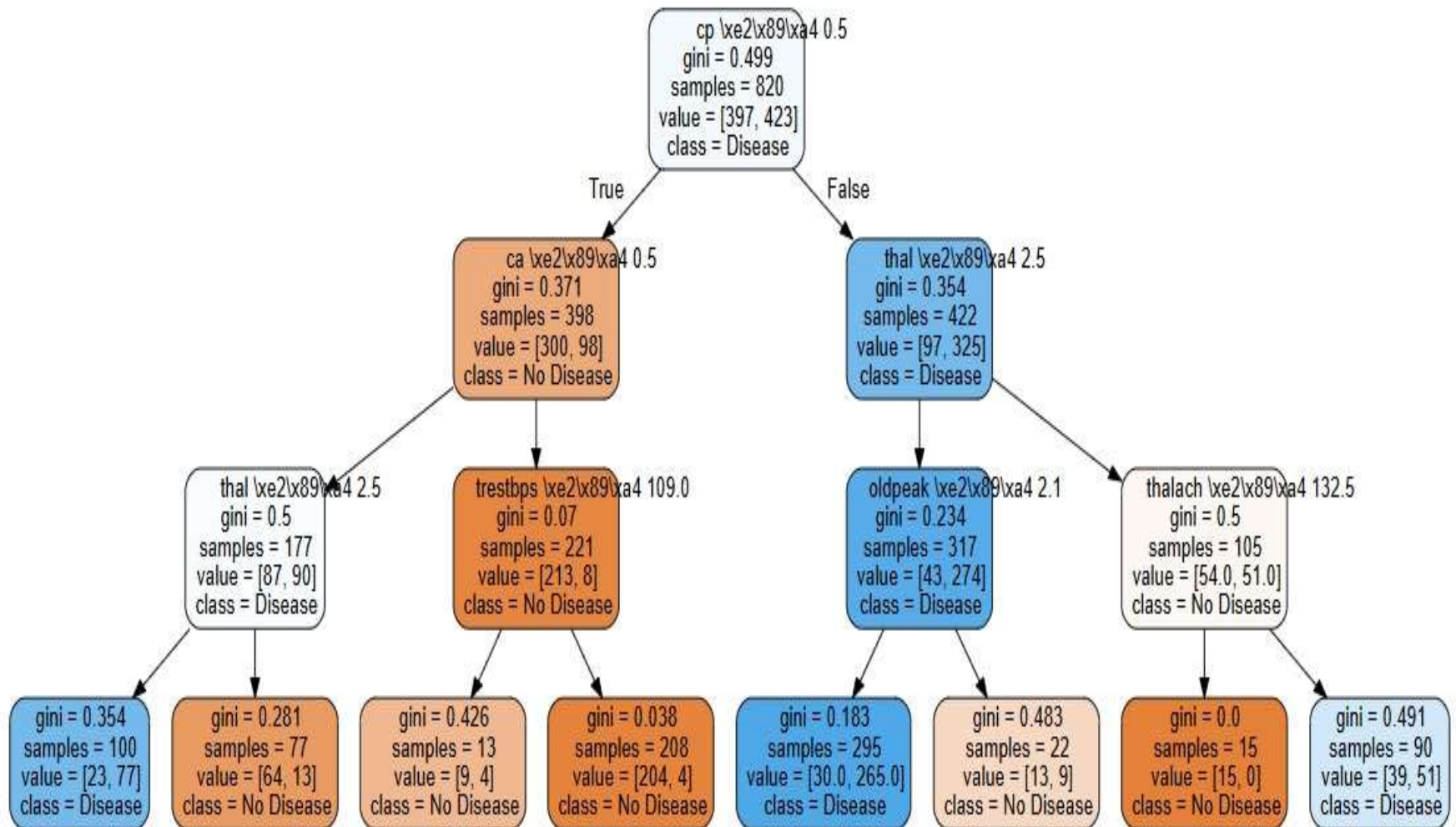**Model trained:** "C:\Users\User\Downloads\Decision Trees and Random Forests.ipynb"

📋 Final Summary:

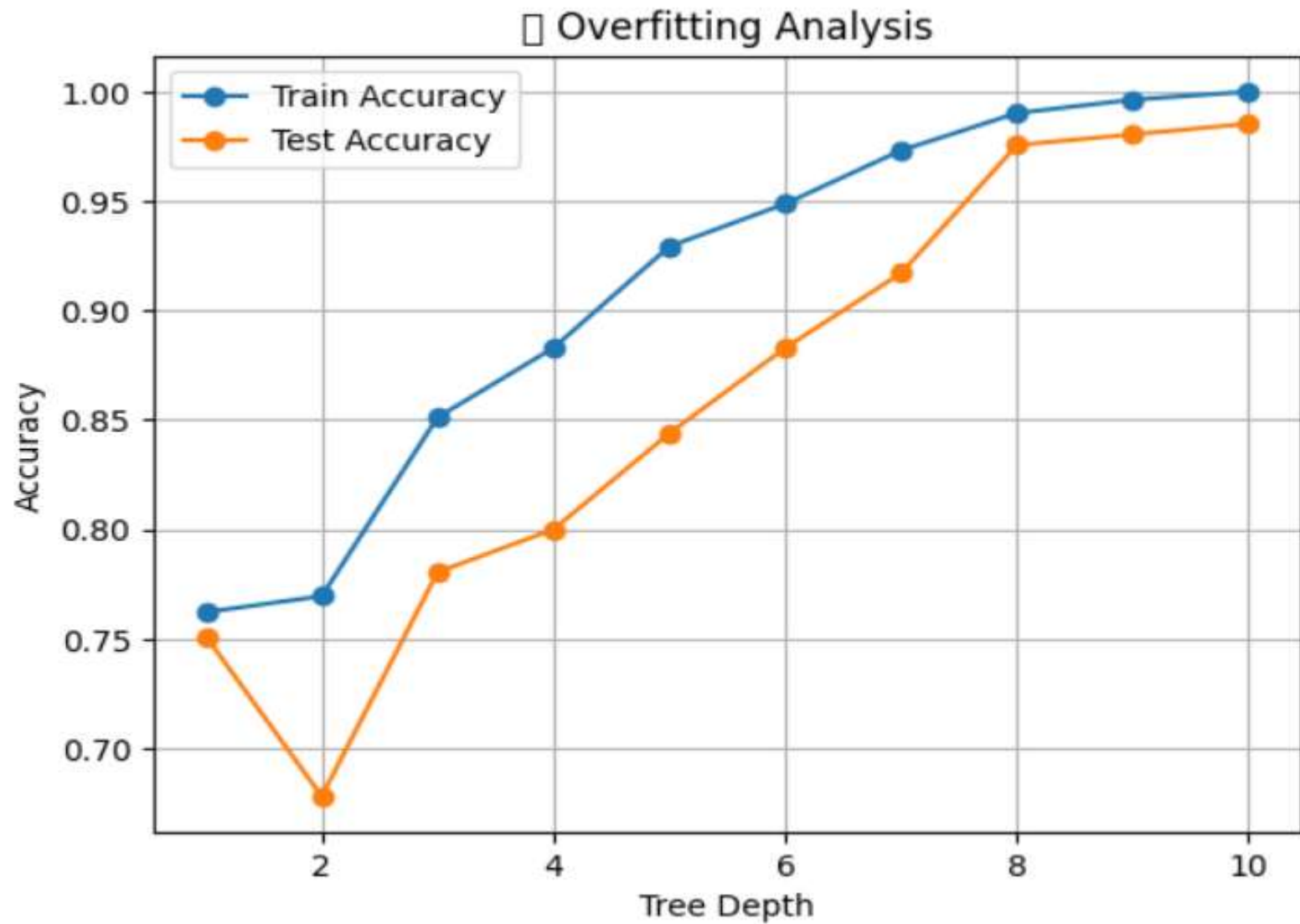🌳 **Decision Tree** - Test Accuracy: 0.7805 | CV Accuracy: 0.8302
🌲 **Random Forest** - Test Accuracy: 0.8732 | CV Accuracy: 0.9298

DECISION TREE VISUALIZATION

cp \xe2\x89\xa4 0.5
gini = 0.499
samples = 820
value = [397, 423]
class = Disease

True — False

ca \xe2\x89\xa4 0.5
gini = 0.371
samples = 398
value = [300, 98]
class = No Disease

thal \xe2\x89\xa4 2.5
gini = 0.354
samples = 422
value = [97, 325]
class = Disease

thal \xe2\x89\xa4 2.5
gini = 0.5
samples = 177
value = [87, 90]
class = Disease

trestbps \xe2\x89\xa4 109.0
gini = 0.07
samples = 221
value = [213, 8]
class = No Disease

oldpeak \xe2\x89\xa4 2.1
gini = 0.234
samples = 317
value = [43, 274]
class = Disease

thalach \xe2\x89\xa4 132.5
gini = 0.5
samples = 105
value = [54.0, 51.0]
class = No Disease

gini = 0.354
samples = 100
value = [23, 77]
class = Disease

gini = 0.281
samples = 77
value = [64, 13]
class = No Disease

gini = 0.426
samples = 13
value = [9, 4]
class = No Disease

gini = 0.038
samples = 208
value = [204, 4]
class = No Disease

gini = 0.183
samples = 295
value = [30.0, 265.0]
class = Disease

gini = 0.483
samples = 22
value = [13, 9]
class = No Disease

gini = 0.0
samples = 15
value = [15, 0]
class = No Disease

gini = 0.491
samples = 90
value = [39, 51]
class = Disease

♣ Decision Tree Test Accuracy: 0.7805

OVERFITTING   ANALYSIS

Feature Importance from Random Forest

Feature Importances (Random Forest)
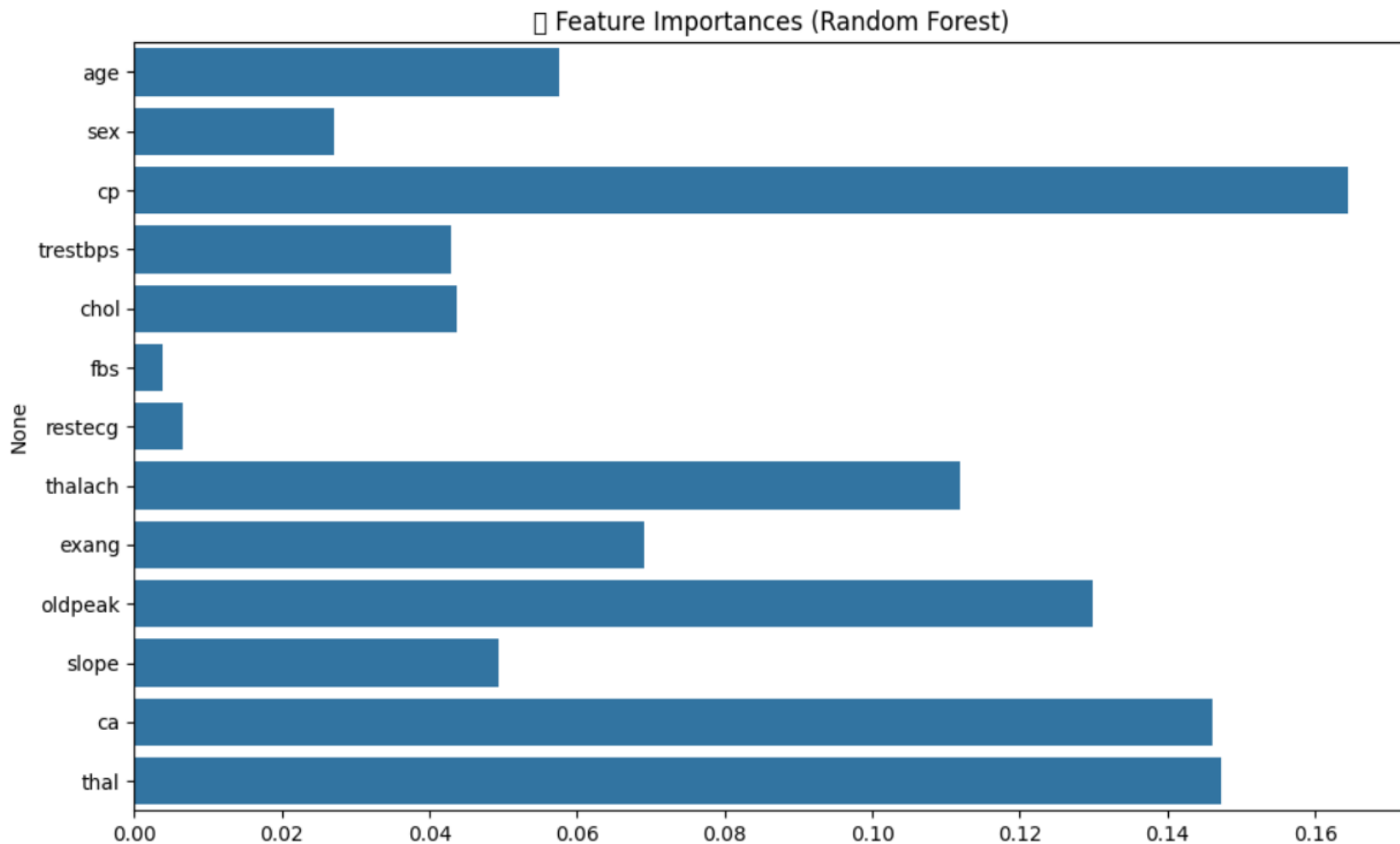
AUTHOR :

S.MADHUMITHA

Github acc : https://github.com/madhu028
linkedin acc : www.linkedin.com/in/madhumitha2810
Mail id : madhumithas2810@gmail.com