# **Heart Disease Classification Analysis**

This project analyses the **heart disease dataset** to predict the presence of heart disease using machine learning models. The analysis includes **data exploration**, **preprocessing**, **model training**, **evaluation**, and **feature interpretation**.

## **1. Dataset Overview**

* Dataset: heart\_database.csv
* Target variable: condition (1 = presence of heart disease, 0 = absence)
* Features: Age, sex, chest pain type, blood pressure, cholesterol, etc.
* Total records: **df.shape[0]**
* Data quality checks:
  + Missing values: None
  + Duplicates: None

**Initial Exploration:**

* Correlation matrix highlights relationships between features.
* Histograms show the distribution of numerical features.

## **2. Preprocessing**

* Removed leading/trailing spaces from column names.
* One-hot encoded categorical variables: cp, restecg, slope, thal.
* Scaled numerical features: age, trestbps, chol, thalach, oldpeak.
* Train-test split: 80% training, 20% testing, stratified by target.

## **3. Model Training**

### **Logistic Regression**

* Trained on scaled and encoded features.

Accuracy score on LogisticRegression model is 91.67%

Confusion Matrix:

[[32 0]

[ 5 23]]

**Classification Report:**

**precision** **recall f1-score support**

0 0.86 1.00 0.93 32

1 1.00 0.82 0.90 28

**accuracy** 0.92 60

**macro avg**  0.93 0.91 0.91 60

**weighted avg** 0.93 0.92 0.92 60

### **Random Forest**

* 100 estimators, trained on the same dataset.

Accuracy score on LogisticRegression model is 91.67%

Confusion Matrix:

[[32 0]

[ 5 23]]

**Classification Report:**

**precision** **recall f1-score support**

0 0.86 1.00 0.93 32

1 1.00 0.82 0.90 28

**accuracy**  0.92 60

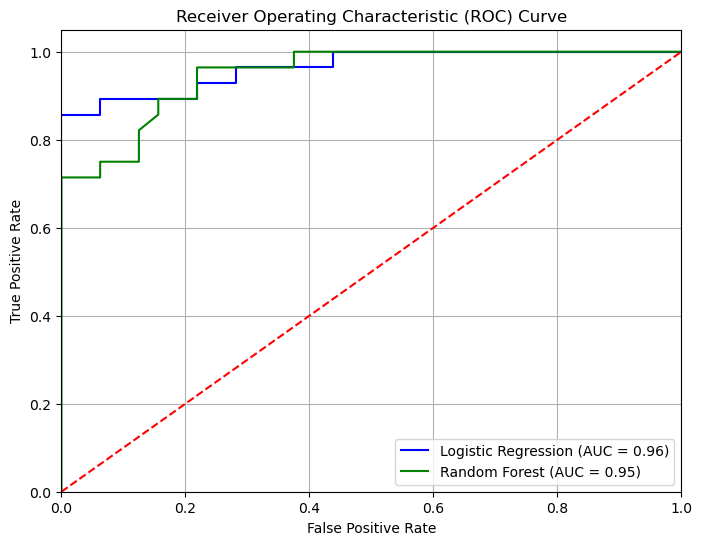
**macro avg** 0.93 0.91 0.91 60

**weighted avg** 0.93 0.92 0.92 60

## **4. Model Evaluation: ROC & AUC**

The **ROC curves** compare both models' ability to distinguish heart disease:

* Logistic Regression AUC
* Random Forest AUC

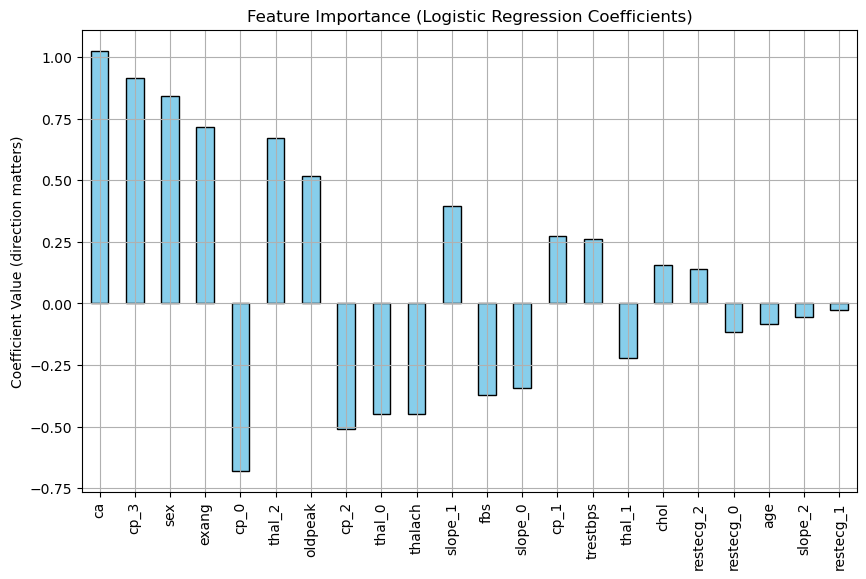


**Observation:**

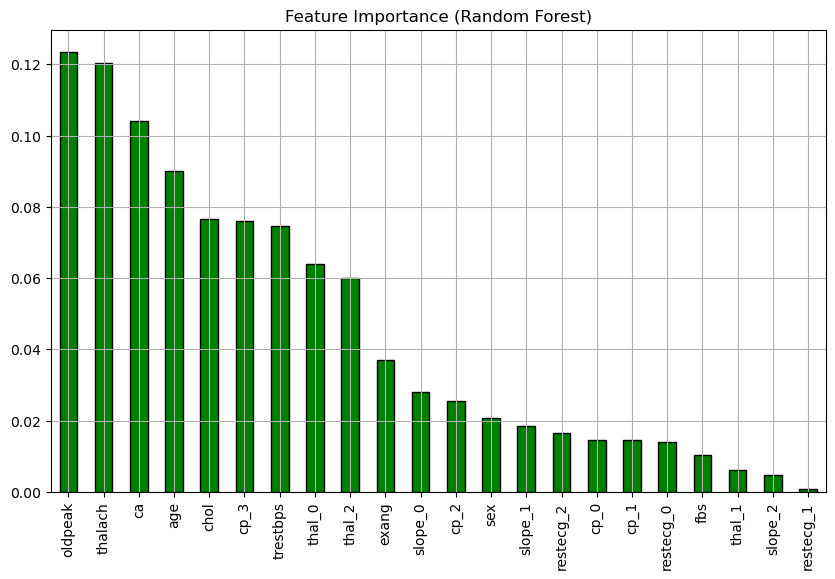
* Both models perform well, with RF slightly outperforming LR.

## **5. Feature Importance**

### **Logistic Regression**

* Feature importance is determined by coefficient magnitude (direction matters).
* **Top 5 features influencing heart disease prediction:**
* Top 5 Logistic Regression features:
  + ca: 1.03
  + cp\_3: 0.92
  + sex: 0.84
  + exang: 0.72
  + cp\_0: -0.68

### **Random Forest**

* Feature importance is determined by Gini importance.
* **Top 5 features influencing heart disease prediction:**  
   

**Observation:**

* Both models highlight similar important features (e.g., cp, thalach, oldpeak).
* RF captures nonlinear relationships better, while LR shows the direction of influence.
* Top 5 Random Forest features:
  + oldpeak: 0.12
  + thalach: 0.12
  + ca: 0.10
  + age: 0.09
  + chol: 0.08

## **6. Conclusion**

* **Random Forest** slightly outperforms Logistic Regression in accuracy and AUC.
* Feature analysis identifies key predictors of heart disease, which can guide medical interpretation.
* The workflow demonstrates the process from **data exploration → preprocessing → model training → evaluation → interpretation**.

Report points:

* No cleaning or sorting was needed for this dataset. The dataset was perfectly created. Clean and full. The only issue i saw while trying to plot some of or use some of the columns was even though it was spelled correctly it will say it doesn’t exist. The issue was merry that there was space in front of the names which made it look like a character to the file. I used **str.strip** to remove any access space.
* Histogram and heatmap to have a visual overview of the dataset.
* For the**reset\_index** i have it as a comment, the reason for that being that while working with this analysis i will be needing to review the original dataset or need to correct the code and it will be giving me the ‘it does not exit’ or similar comment. To get rid of that with no issue i use this code to rest and start without needing to run the whole code again and restarting.
* Using condition as the target before starting with the scaling and encoding and dropping from X. Why conditions. Since we are trying to see if we can build a ml model that will be able to predict who of the patient is most likely to suffer from a heart disease and using condition as the target since it already has the list of who does and doesn’t.