Phase 3 – Feature Selection Report

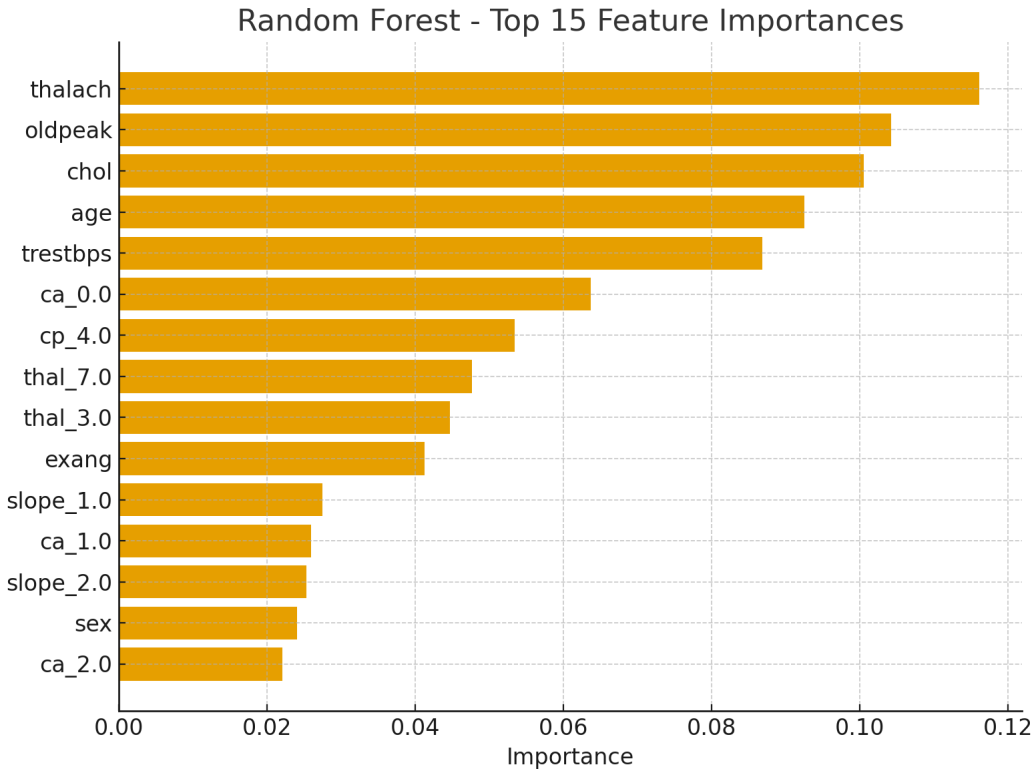
The goal of Phase 3 is to reduce the dimensionality of the dataset by selecting the most relevant features that have the highest impact on predicting heart disease. This helps to avoid overfitting, improve model interpretability, and speed up training.

Methods Used:
1. Random Forest Feature Importance – used to rank the features by importance.
2. (Optional) Other methods like RFE and Chi-Square can be applied, but the primary method here is Random Forest.

Results:
The top features identified by Random Forest were mainly related to heart rate, ST depression, cholesterol, and age. These are considered key indicators for heart disease prediction.

Rank	Feature (Definition)
1	thalach – maximum heart rate achieved
2	oldpeak – ST depression induced by exercise relative to rest
3	chol – serum cholesterol in mg/dl
4	age – age of the patient
5	trestbps – resting blood pressure in mm Hg
6	ca_0.0 – number of major vessels (value 0, one-hot encoded)
7	cp_4.0 – chest pain type (category 4, encoded)
8	thal_7.0 – thalassemia (category 7, encoded)
9	thal_3.0 – thalassemia (category 3, encoded)
10	exang – exercise-induced angina (1 = yes, 0 = no)

Feature Importance Visualization:



Conclusion:
From the feature selection process, the dataset was reduced to the most important variables. A new dataset **selected_features.csv** was created, containing only these features and the target. This reduced dataset will be used in Phase 4 (Supervised Learning) and Phase 5 (Unsupervised Learning).