Phase 3 – Feature Selection Report

The goal of this phase is to identify the most important features that contribute to heart disease prediction. Feature selection helps reduce unnecessary variables, improve model performance, and make the model easier to understand.

Method Used: Random Forest Feature Importance was used to rank features based on how much they contribute to the prediction.

Random Forest is suitable because:

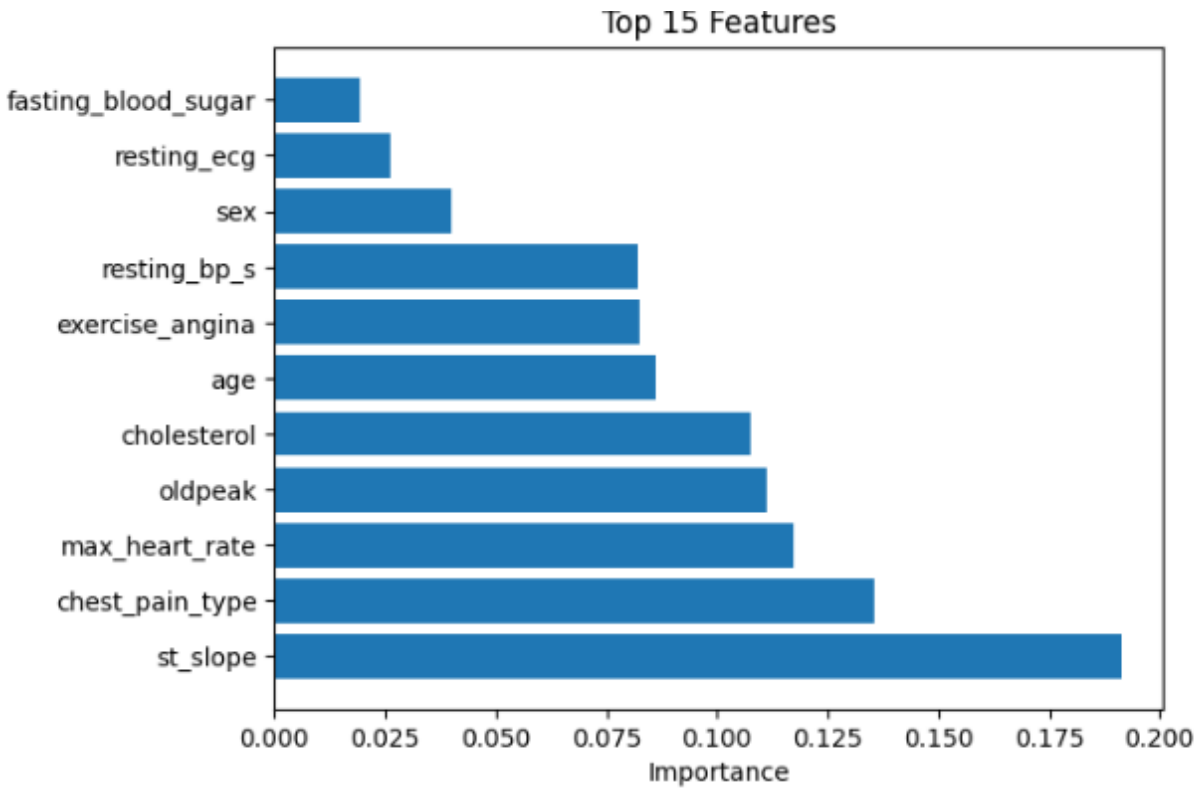
- It handles non-linear relationships
- It provides built-in feature importance
- It works well with medical tabular data

Results:

	feature	importance
0	st_slope	0.191213
1	chest_pain_type	0.135547
2	max_heart_rate	0.117334
3	oldpeak	0.111069
4	cholesterol	0.107771
5	age	0.086331
6	exercise_angina	0.082646
7	resting_bp_s	0.082216
8	sex	0.040066
9	resting_ecg	0.026394
10	fasting_blood_sugar	0.019414

Feature Importance Visualization

A bar chart was created to visualize the top 15 features ranked by importance.
The plot clearly shows that **ST slope**, **chest pain type**, and **maximum heart rate** are the most influential features.



Conclusion

- Feature selection helped reduce the dataset to the most relevant variables.
- A new dataset named **`selected_features.csv`** was created containing only important features and the target variable.
- This dataset was used for model training and hyperparameter tuning in later phases.