Yeditepe University
Department of Computer Engineering
Data Science Project

Heart Disease Prediction Using Machine Learning

Berkay Kuru (20200702115)
2025

Heart Disease Prediction Using Machine Learning

1) Introduction

Heart disease remains one of the leading causes of mortality worldwide. Early detection is essential for preventing severe outcomes, yet traditional diagnostic processes often rely on manual evaluation and subjective interpretation.
With the increasing availability of structured medical data and the rise of machine learning (ML), data-driven prediction has become a promising approach for clinical decision support.This study aims to build predictive models that estimate whether a patient has heart disease based on clinical attributes such as chest pain type, cholesterol level, resting blood pressure, maximum heart rate, and ST depression induced by exercise.
Using Logistic Regression and Random Forest algorithms, the project evaluates model accuracy, precision, recall, F1-score, and ROC-AUC metrics to determine which approach provides the best balance between interpretability and predictive power.

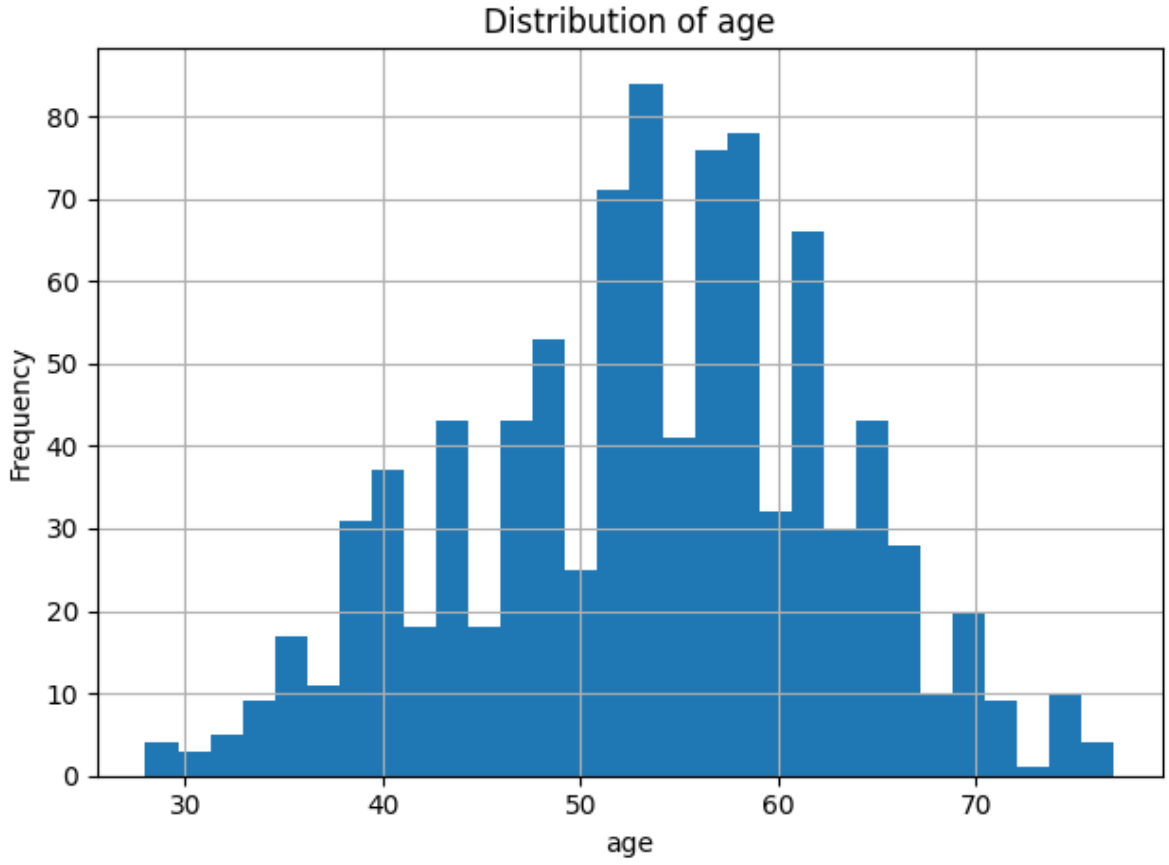
2) Data Understanding

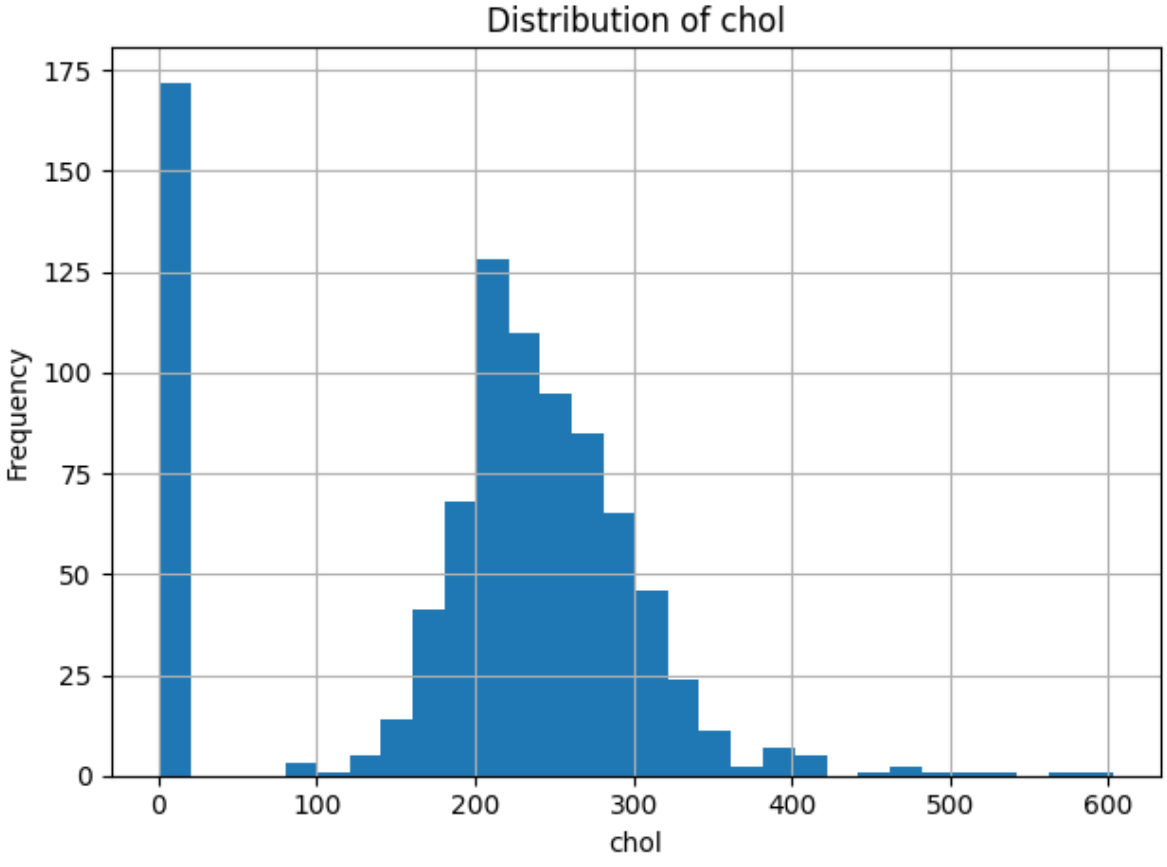
The dataset used in this project originates from the UCI Heart Disease Repository, a benchmark dataset widely used in health-related data analysis. It contains 303 patient records described by 14 features and one target label.

2.1) Feature Description

Feature	Description
age	Patient's age in years
gender	1 = male, 0 = female
cp	Chest pain type (typical angina, atypical angina, non-anginal pain, asymptomatic)
trestbps	Resting blood pressure (mmHg)
chol	Serum cholesterol level (mg/dl)
thalach	Maximum heart rate achieved
oldpeak	ST depression induced by exercise relative to rest
target	1 = presence of heart disease, 0 = absence

The target variable ("target") indicates whether heart disease is present or not.
Most patients fall into the 40–65 age range, and roughly 54% of the records correspond to positive heart disease cases.


Distribution of age
Frequency
0
10
20
30
40
50
60
70
80
30
40
50
60
70
age

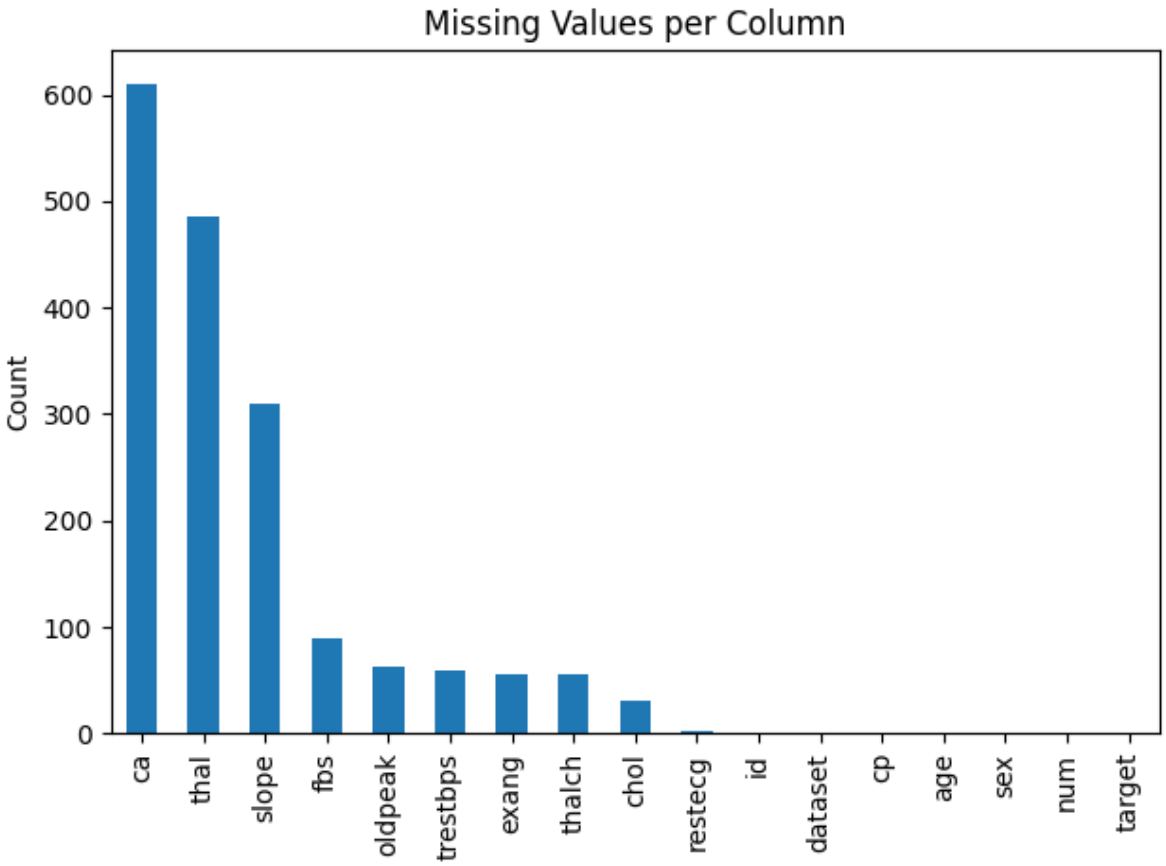

Distribution of chol
Frequency
0
25
50
75
100
125
150
175
0
100
200
300
400
500
600
chol

3) Data Cleaning and Preprocessing

Prior to model training, the dataset was cleaned and prepared to ensure consistency and quality.

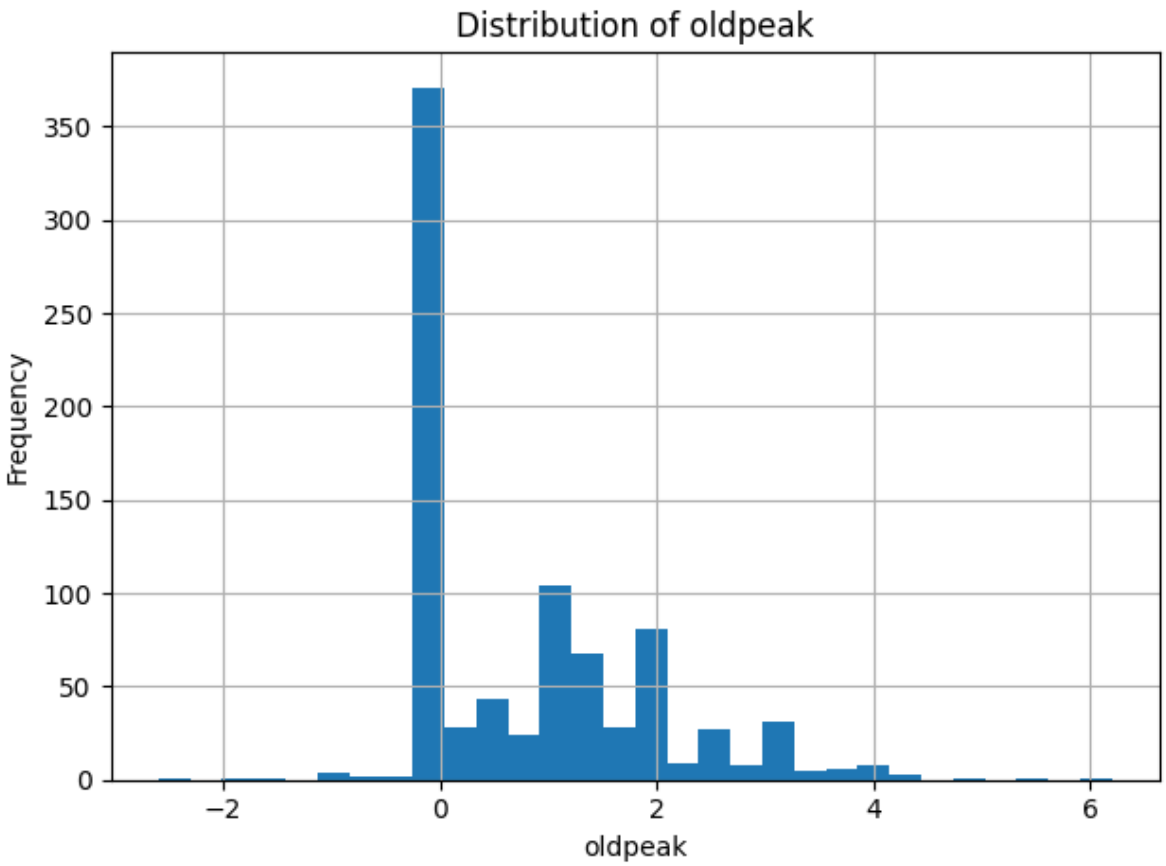
3.1) Missing Values

An analysis of missing values showed several incomplete entries, primarily in the ca, thal, and slope columns.
Categorical features were imputed with their most frequent values, while continuous features (e.g., chol, trestbps) were filled with their median.



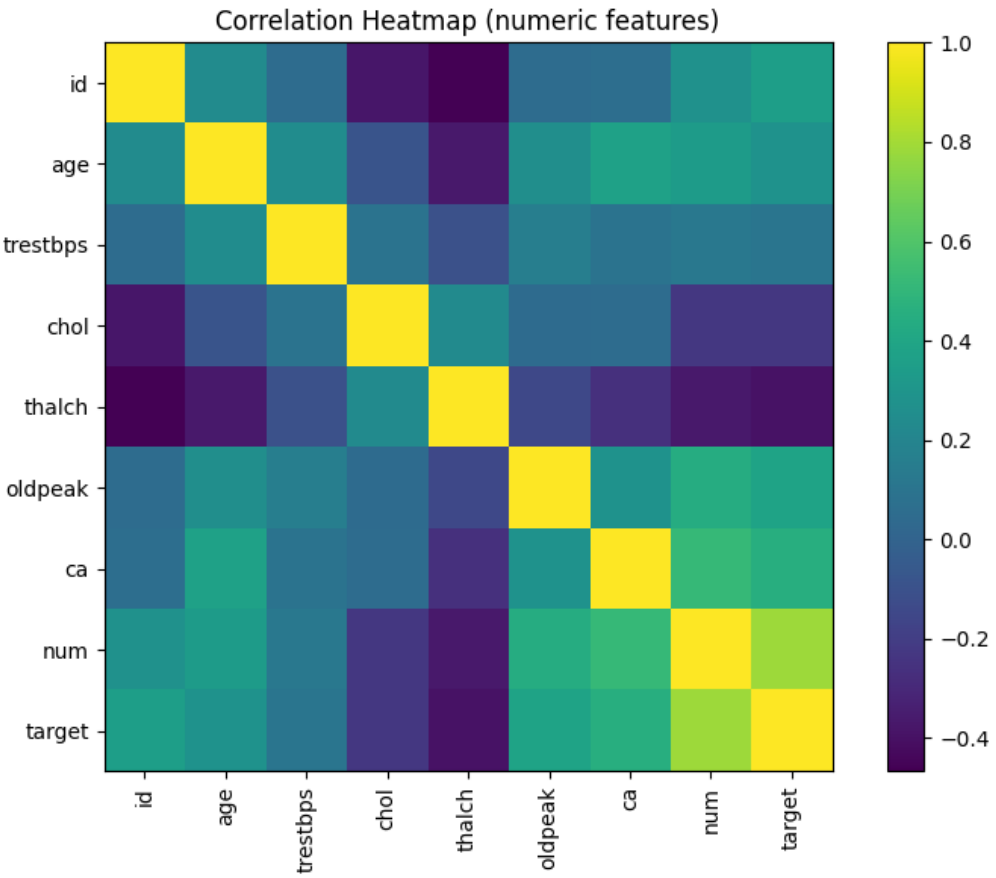
3.2) Outlier Handling

Visual inspection through histograms indicated that most numerical features follow near-normal distributions, with a few high-end outliers in cholesterol and oldpeak. These outliers were retained, as they represent real-world variability in clinical data.



3.3) Feature Scaling and Encoding

Continuous variables were standardized using Z-score normalization, while categorical variables (such as cp, slope, thal) were transformed via One-Hot Encoding.



3.4) Dataset Splitting

After preprocessing, the dataset was divided into:

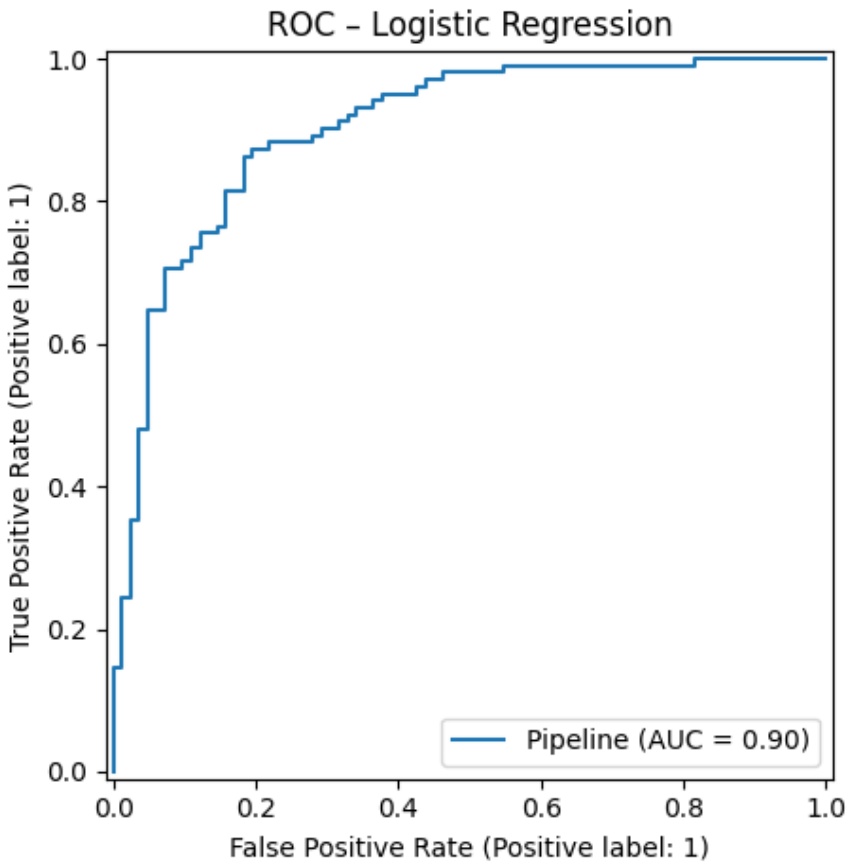
- Training set: 80% (242 samples)
- Testing set: 20% (61 samples)

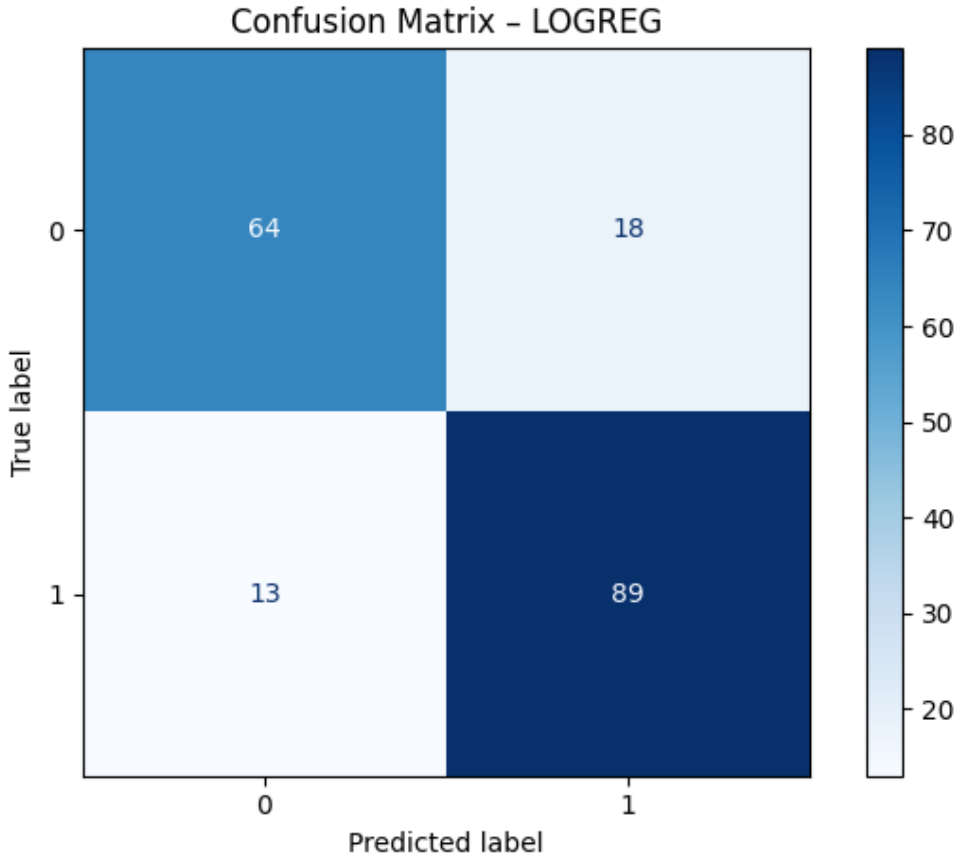
4) Dataset Splitting

Two machine learning algorithms were implemented and evaluated: Logistic Regression (a linear baseline) and Random Forest (a non-linear ensemble method). Model training, tuning, and validation were performed using a 3-fold Stratified Cross-Validation scheme.

4.1) Logistic Regression

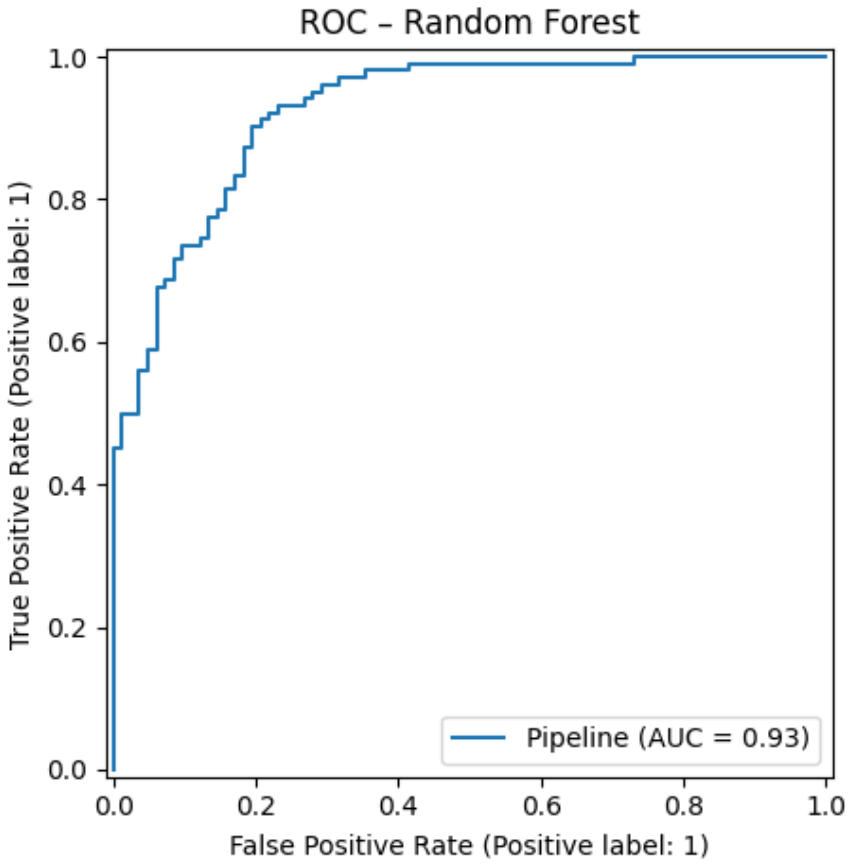
Logistic Regression was chosen for its interpretability and efficiency in binary classification.
The optimal hyperparameters (C=0.1, penalty=L2) were found via grid search.
This model achieved an accuracy of 83.2% and F1-score of 85.2% on the test set.

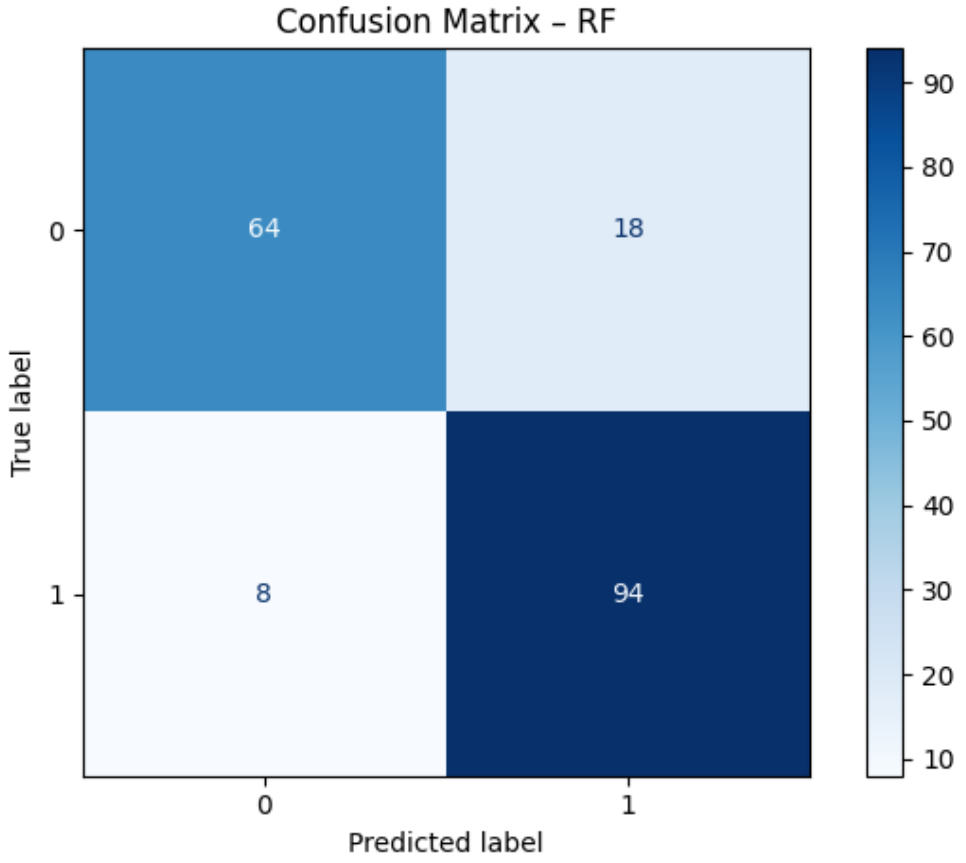



Confusion Matrix – LOGREG
True label
Predicted label
0
1
0
1
64
18
13
89
80
70
60
50
40
30
20

4.2) Random Forest

Random Forest, an ensemble of decision trees, was used to capture non-linear interactions among features.
With 200 trees and max depth tuned to 8, it achieved an accuracy of 85.9%, F1-score of 87.9%, and the highest ROC-AUC of 0.92 among all models.





4.3) Model Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.832	0.832	0.873	0.852	0.88
Random Forest	0.859	0.839	0.922	0.879	0.92

Random Forest outperformed Logistic Regression in all major evaluation metrics, showing superior recall and robustness against overfitting.

5) Feature Importance and Insights

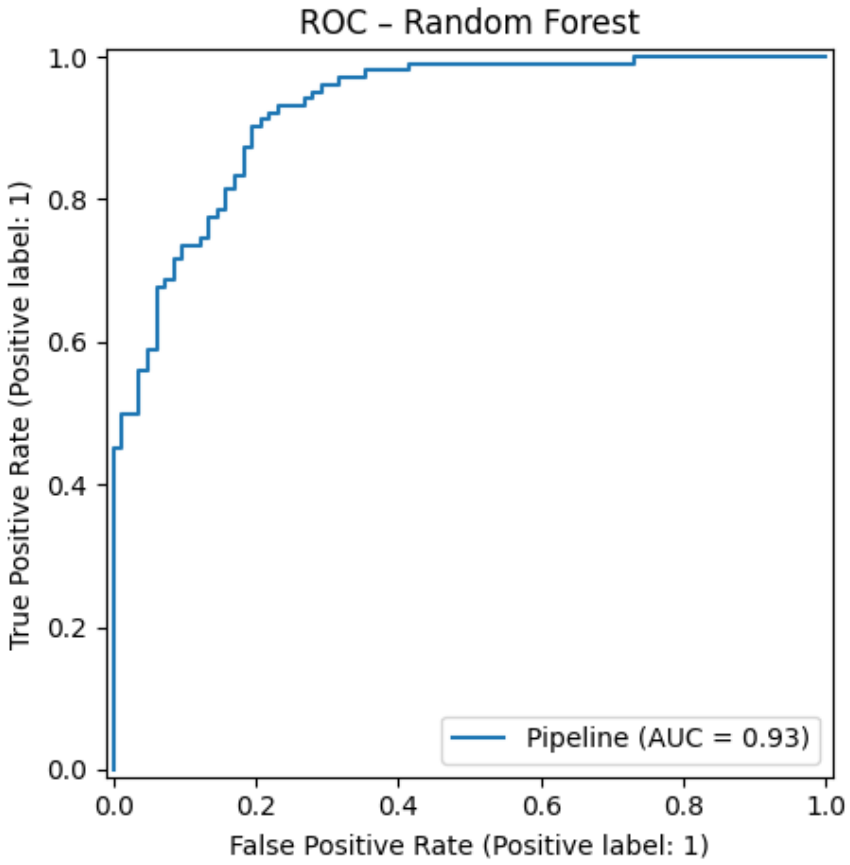
To interpret the Random Forest model, the top 10 most influential features were extracted.

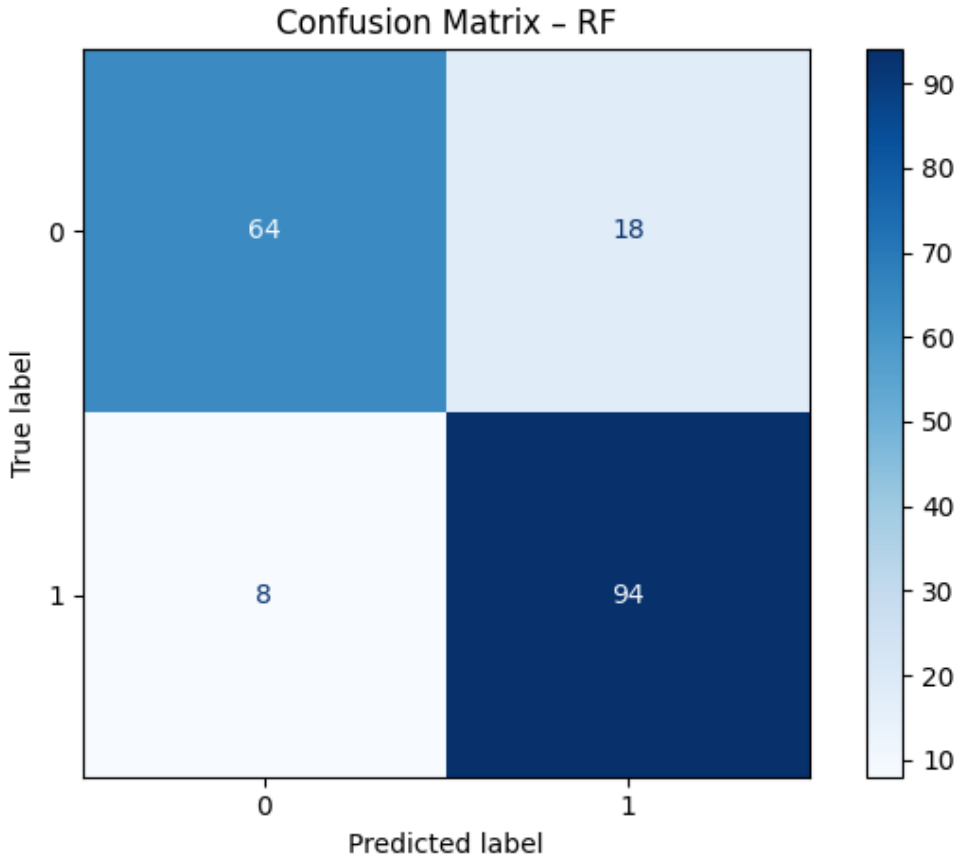
Rank	Feature	Importance
1	cp_asymptomatic	0.142
2	chol	0.102
3	oldpeak	0.096
4	age	0.095
5	thalach	0.093
6	exang_False	0.076
7	exang_True	0.065
8	cp_atypical angina	0.056
9	trestbps	0.045
10	sex_Male	0.033

The most impactful factors were chest pain type (cp_asymptomatic), cholesterol, oldpeak, and age, which aligns with clinical expectations in cardiology.
Higher cholesterol and lower maximum heart rate tend to increase the predicted probability of heart disease.

6) Results and Discussion

Overall, both models provided strong predictive performance. The Random Forest model offered a better balance between precision and recall, making it suitable for screening scenarios where identifying positive cases (true positives) is more critical than minimizing false alarms.





The interpretability analysis also confirmed that the model learned clinically meaningful patterns, reinforcing its reliability for practical applications.

7) Conclusion

This study successfully applied machine learning techniques to predict heart disease using clinical data.
Among the models tested, the Random Forest classifier achieved the best overall performance with an accuracy of 85.9% and an AUC of 0.92.
Feature importance analysis revealed that chest pain type, cholesterol, age, and ST depression are the key predictive factors.

Future work could include:

- Applying XGBoost or LightGBM for further performance improvement.
- Integrating patient data from multiple sources to increase generalizability.
- Deploying the model via a web or mobile interface for real-time risk estimation.

8) References

- UCI Machine Learning Repository: Heart Disease Dataset
- Han, J. & Kamber, M. (2012). Data Mining: Concepts and Techniques
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning
- Scikit-learn documentation: https://scikit-learn.org