Midpoint Report — Heart Disease (HeartDisease_G2)

Updated Dataset Description & Cleaning:

• UCI Cleveland Heart Disease (303 rows, 14 attributes).

• target = 1 if disease present (num>0), else 0; regression target = thalach (bpm).

• ca (4 rows) & thal (2 rows) imputed by mode; standardization & one-hot via sklearn Pipeline.

EDA:

• Slight class imbalance (~55% positive).

• thalach inversely related to age; oldpeak tends to increase with exang.

Split & Baselines:

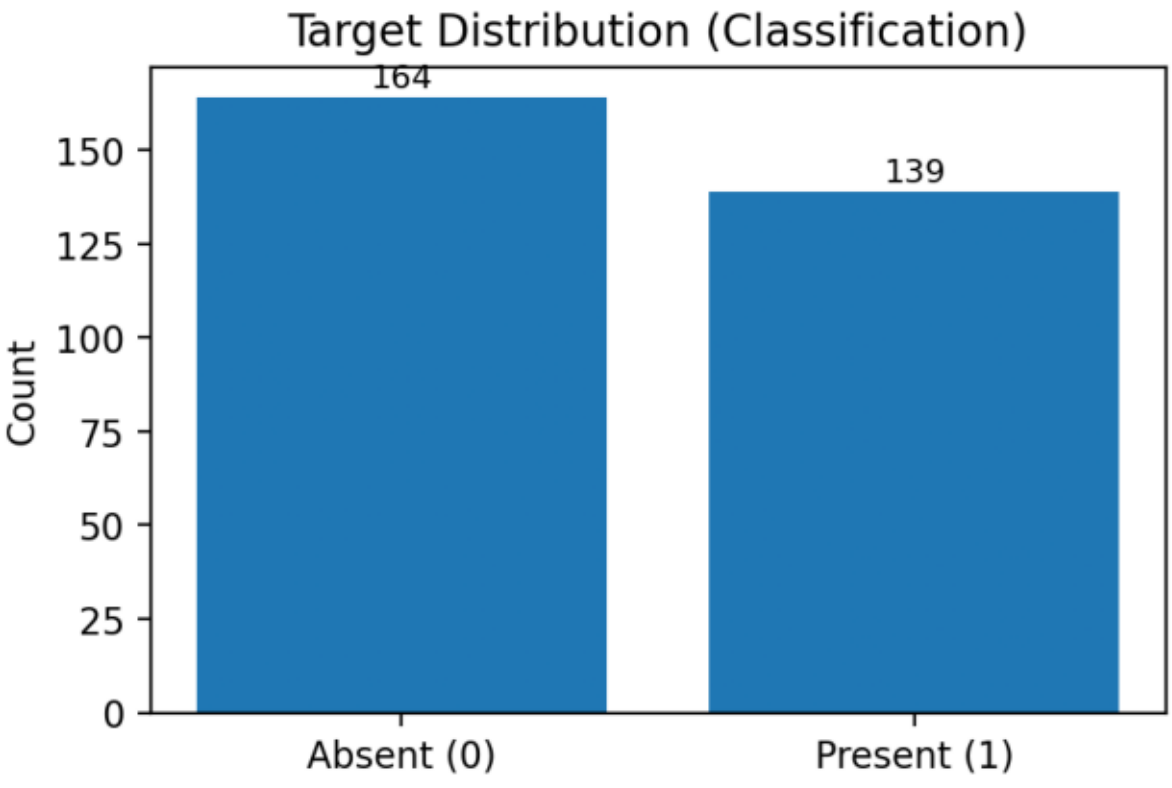
• 60/20/20 train/val/test split, random_state=42; stratified for classification.

• Classification baselines: Logistic Regression, Decision Tree (MLflow tracked).

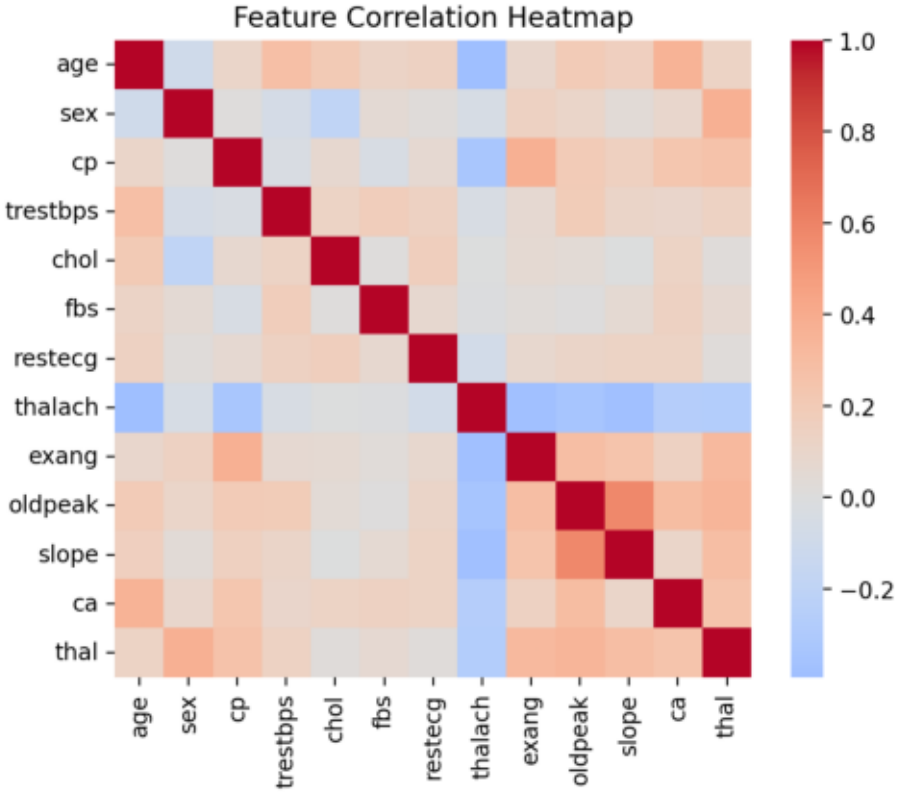
• Regression baselines: Linear Regression, Decision Tree Regressor (MLflow tracked).

Plots on following pages match rubric exactly: 1) target distribution, 2) correlation heatmap, 3) confusion matrix of best classification baseline on TEST, 4) residuals vs predicted of best regression baseline on TEST.

Plot 1 — Target Distribution



Plot 2 — Correlation Heatmap



Plot 3 — Confusion Matrix (Test)

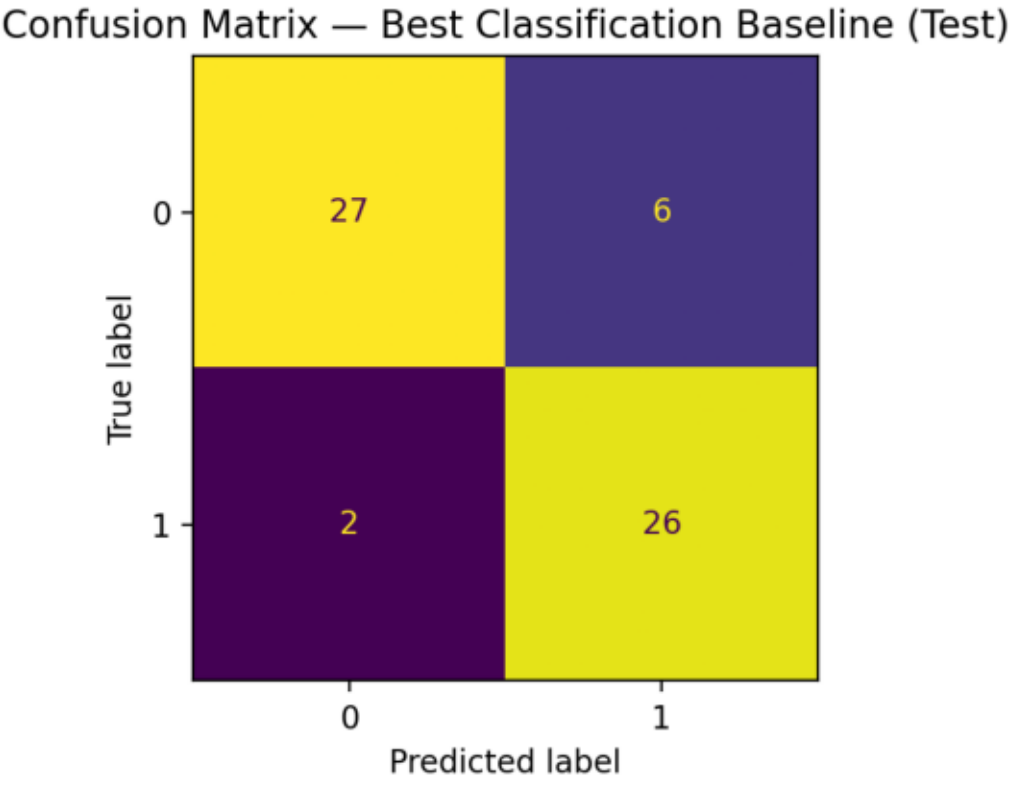


Table 1 — Classification Metrics (Val & Test)

Model	Accuracy_val	F1_val	Accuracy_test	F1_test
LogisticRegression	0.8197	0.8179	0.8689	0.8688
DecisionTreeClassifier	0.7213	0.7213	0.7213	0.721

Plot 4 — Residuals vs Predicted (Test)

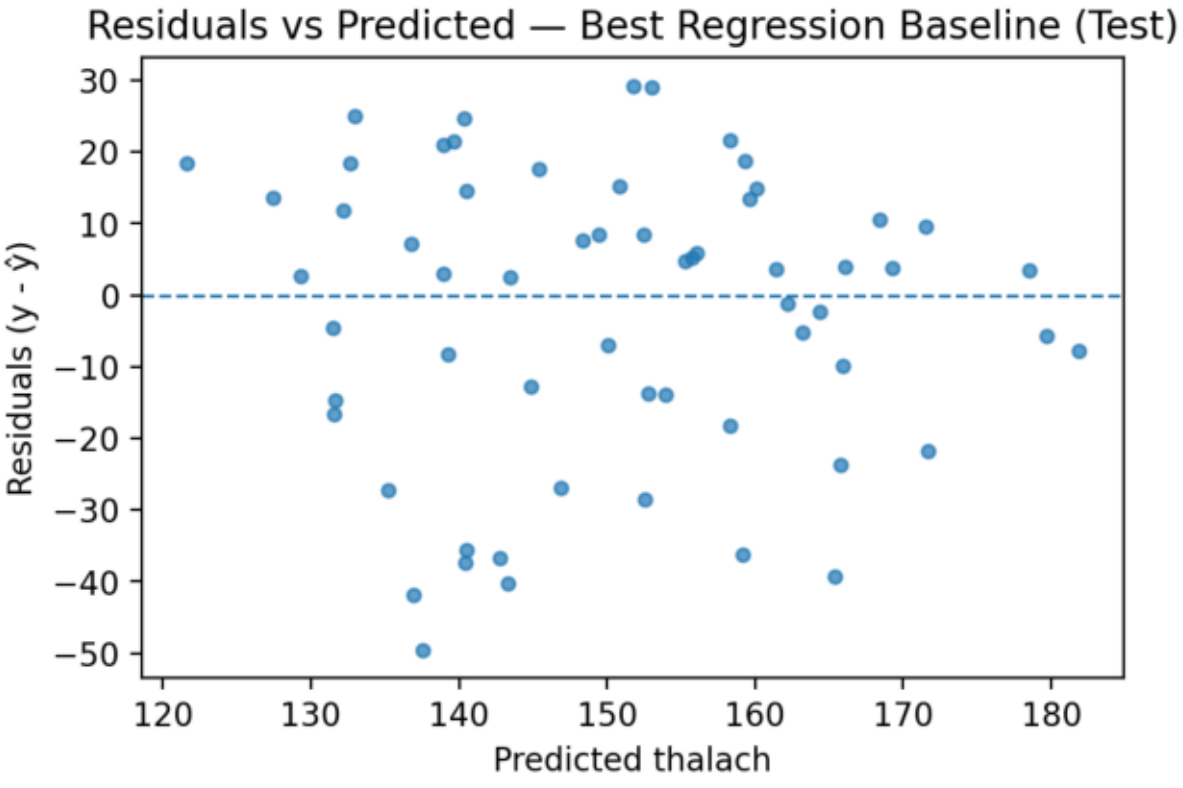


Table 2 — Regression Metrics (Val & Test)

Model	MAE_val	RMSE_val	MAE_test	RMSE_test
LinearRegression	12.5512	15.468	16.5636	20.2138
DecisionTreeRegressor	16.377	22.0115	21.5902	27.187

Results & Discussion (brief):

• The best classification baseline is chosen by highest validation F1; we report test confusion matr

• The best regression baseline is chosen by lowest validation RMSE; residual plot shows remaining

• Typical errors: borderline oldpeak/exang combinations (FPs); consider regularization/depth tunin

Neural Network Plan:

• Two MLPs (classification & regression): 64→32→16 (ReLU), BatchNorm + Dropout 0.2.

• Adam (lr=1e-3), batch 64, early stopping on val F1 (cls) and val RMSE (reg).

• Standardize numeric, one-hot categorical; log learning curves and final metrics with MLflow.