Kevin O'Donnell
June 1st, 2025

1. INTRO/DATASET

Heart disease is the leading cause of death worldwide, responsible for nearly 18 million deaths annually. Supervised learning offers a powerful approach to accelerate early detection by analyzing clinical data–saving time, reducing healthcare costs, and improving patient outcomes. This project focuses on predicting the presence or absence of cardiovascular disease using supervised binary classification. The Heart Failure Prediction dataset was created by compiling five existing heart disease sources: Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog. After cleaning and integration, the final dataset contains 918 patients and 11 clinical features commonly associated with heart disease risk.

1) Age
2) Sex
3) Chest Pain Type
    a) TA: Typical Angina
    b) ATA: Atypical Angina
    c) NAP: Non-Anginal Pain
    d) ASY: Asymptomatic
4) Resting Blood Pressure (mmHg)
5) Cholesterol Level (mm/dl)
6) Fasting Blood Sugar (1 if > 120 mg/dl, else 0)
7) Resting ECG
    a) Normal
    b) ST = ST-T wave abnormality
    c) LVH = Left Ventricular Hypertrophy
8) Max Heart Rate (range 60-202)
9) Exercise Angina (Y or N)
10) Oldpeak: ST depression induced by exercise relative to rest
11) ST Slope: slope of the peak exercise ST segment
    a) Up = upsloping
    b) Flat = flat
    c) Down = downsloping
12) Heart Disease = TARGET Y
    a) 1 = presence of heart disease
    b) 0 = no heart disease

2. PREPROCESSING

The most important step in supervised learning is ensuring the dataset is both clean and physiologically plausible. The dataset did not contain any missing (NaN) values, however, several features included medically implausible entries. Specifically, rows with Resting Blood Pressure (RestingBP) or Maximum Heart Rate (MaxHR) equal to zero were

removed. Similarly, any negative values in Oldpeak were removed. This cleaning step slightly reduced the dataset, from 918 to 904 patients, and was preferred over other value replacement strategies to avoid introducing unrealistic values. To mitigate the influence of outliers without discarding additional patients, Cholesterol values were clipped at an upper limit of 400 mg/dL, capping extremely high values while preserving the data.

Following this, the dataset was split into three mutually exclusive subsets: training (65%), validation (20%), and test (15%). Stratified sampling ensured that the proportion of positive vs. negative heart disease cases remained consistent across all subsets, while a random state of 42 was used to maintain reproducibility.

A challenge arose with the Cholesterol feature, where 159 patients had a recorded value of 0 mg/dL, which is clinically invalid. Rather than dropping a large portion of the dataset, an Iterative Imputer was chosen as an informed approach to replace these invalid values. This imputation strategy provided a more physiologically meaningful alternative to using simple mean or median imputation. The scikit-learn's IterativeImputer method was set up to predict cholesterol values for each patient that had a 0 value based on the continuous features: Age, RestingBP, MaxHR, and Oldpeak. The imputation was performed following the train-test splits, so the Imputer was only trained on the training set to prevent data leakage. The Imputer was then applied to all split sets. Due to any introduced variance in this step, despite taking a well-suited approach to filling in these physiologically implausible values, it will be important later to confirm the model isn't heavily relying on the Cholesterol feature for its decision making.

Preprocessing was then finished with categorical encoding and feature scaling. Binary features (Sex, FastingBS, ExerciseAngina) were mapped using a custom function that applied consistent forward and reverse label encoding (e.g. female -> 0, male -> 1), allowing for later interpretation. Multiclass categorical features (ChestPainType, RestingECG, ST_Slope) were one-hot encoded, converting each category into a separate binary feature. Continuous features (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) were standardized using z-score normalization with scikit-learn's StandardScaler, ensuring models sensitive to feature scale could perform optimally. The training, validation, and test sets were saved as CSV files, while the trained imputer, scaler, and reverse mapping dictionary were saved using joblib, ensuring full reproducibility.

3. INITIAL MODEL SELECTION

To identify high-performing algorithms suitable for this dataset, a preliminary model benchmarking phase was conducted using PyCaret's AutoML functionality. This allowed for quick comparison across an array of classification models ranked on F1 score. F1 score was chosen as the metric because its balance of precision and recall helps ensure both false positives and negatives are minimized. Depending on the nature of the target, a more specific strategy can be implemented (prioritizing low false positives for high-risk or low false negatives if testing is high cost). The top-performing models on this dataset were: Logistic Regression, K-Nearest Neighbors, Gradient Boosting Classifier, and XGBoost.

Kevin O'Donnell
June 1st, 2025

4. HYPERPARAMETER OPTIMIZATION

Hyperparameter optimization was performed for each model using Optuna, which utilizes Bayesian optimization to hone in on the optimal hyperparameters. This is a more targeted and efficient approach than traditional Grid Search or Random Search methods. The hyperparameter search aimed to maximize ROC AUC, a robust metric for initially evaluating performance across all classification thresholds, and utilized stratified k-fold cross-validation. The best hyperparameters for each model were found to be:

Model	AUC Score	Key Hyperparameters
Logistic Regression	0.9268	solver='lbfgs', penalty='l2', C=0.1869
K-Nearest Neighbors	0.9167	n_neighbors=15
Gradient Boosting	0.9444	max_depth=4, learning_rate=0.0213, n_estimators=187, subsample=0.924, min_samples_split=6, min_samples_leaf=10
XGBoost	0.9426	learning_rate=0.0273, max_depth=3, n_estimators=255, subsample=0.581

5. SUBGROUP AUDITING AND FAIRNESS-WEIGHTED RETRAINING

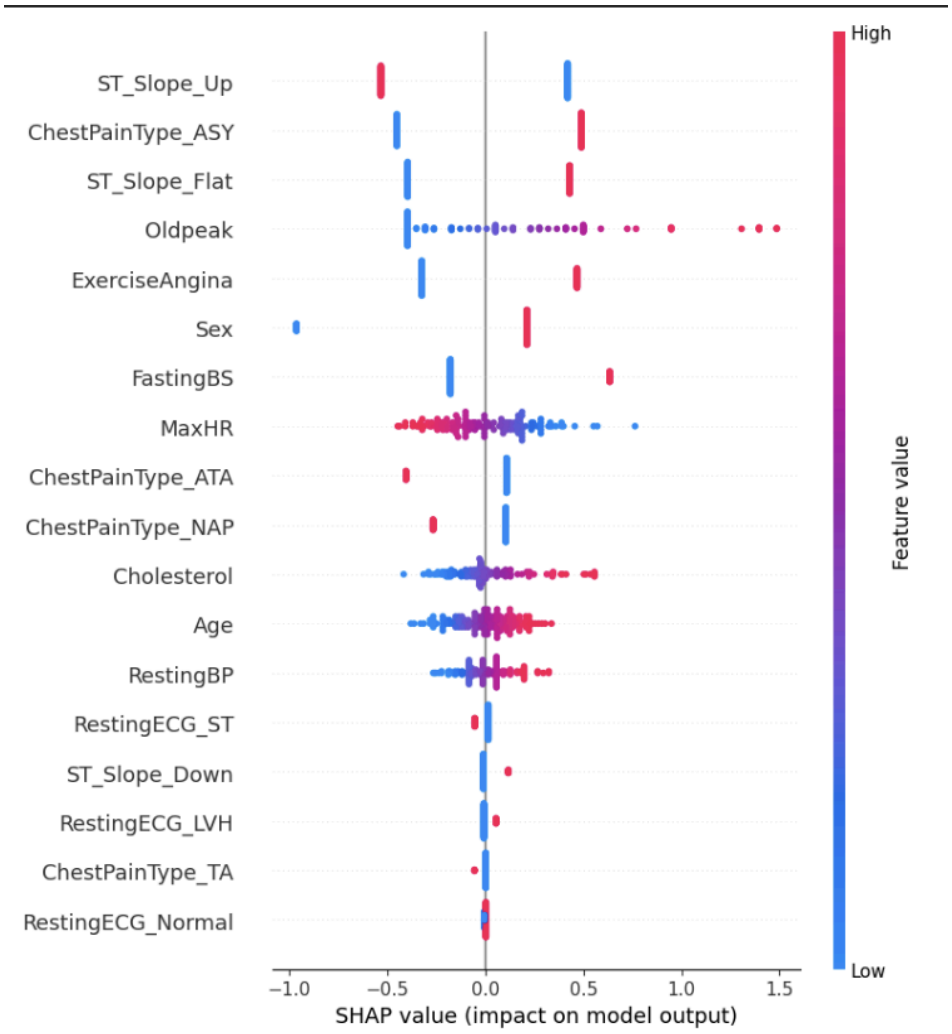
After each model's hyperparameters were optimized on the training set, subgroup auditing was performed to identify where the model was underperforming on the validation set. The validation set served as an intermediary evaluation set to support tuning efforts while preserving the test set for final performance evaluation. A custom subgroup audit function was written, flagging underperforming groups when AUC or F1 dropped below 0.70, or when the Brier score exceeded 0.15. An example of the full custom audit read-out can be found at the bottom of this report. For example, in the Gradient Boosting Classifier (GBC) model, the female subgroup showed considerably lower performance than the male subgroup. This imbalance could result in biased outcomes if the model were deployed without mitigation.

--- GBC Sex_group Unweighted ---

Value	N	AUC	F1 Score	Brier Score	Flags
F	43	0.8374	0.5833	0.1664	High Brier, Low F1
M	138	0.8676	0.8701	0.1281	OK

Kevin O'Donnell
June 1st, 2025

To address this, a fairness-weighted retraining strategy was explored. Specifically, higher weights were assigned to underperforming subgroups—in this case, females—during training to help the model better generalize to these populations. Optuna was used to optimize the weight applied to the subgroup over a search range of 1.0 to 5.0. Weights greater than 1 were chosen deliberately to increase the model's focus on the specified subgroup. The optimization aimed to balance subgroup fairness and overall performance using a weighted metric: 0.2 * subgroup score + 0.8 * overall score. The optimal weight was then applied alongside the previously determined best hyperparameters to retrain the model. The overall and subgroup results after weighting the GBC model are found below.



--- Gbc Classifier Parameters (Weighted) ---
Validation AUC Score (with chosen weights): 0.8758
Validation F1 Score (with chosen weights): 0.8317
Chosen Subgroup Weights: {'Sex_group__F': 1.1542488411525558}

--- GBC Sex_group Weighted ---

Value	N	AUC	F1 Score	Brier Score	Flags
F	43	0.822660	0.300000	0.276179	High Brier, Low F1
M	138	0.867397	0.868571	0.146255	OK

After retraining the GBC model with the chosen weight (1.1542 for females), model performance declined across both the overall and subgroup levels. Despite assigning 15% more attention to the female subgroup, its validation F1 score dropped, and the Brier score worsened. This degradation likely stems from the limited number of female samples, leading the model to overfit to the training data rather than learning meaningful patterns. SHAP analysis in the graph further revealed that the model tended to predict females as negative cases, reinforcing the idea that excessive weighting distorted the model's generalization and increased bias.

Given these results, the fairness-weighted version of the GBC model was discarded in favor of the original, unweighted model, which maintained higher overall performance. The unweighted model was used in all cases since, across the board, applying subgroup-specific weights often reduced global performance without delivering meaningful gains in subgroup fairness. Even attempts to re-tune hyperparameters under the new weights failed to resolve the underlying issue: a low n subgroup with few positive samples. As a result, adjusting subgroup-specific decision thresholds will be adopted later on in the pipeline as the alternative fairness strategy.

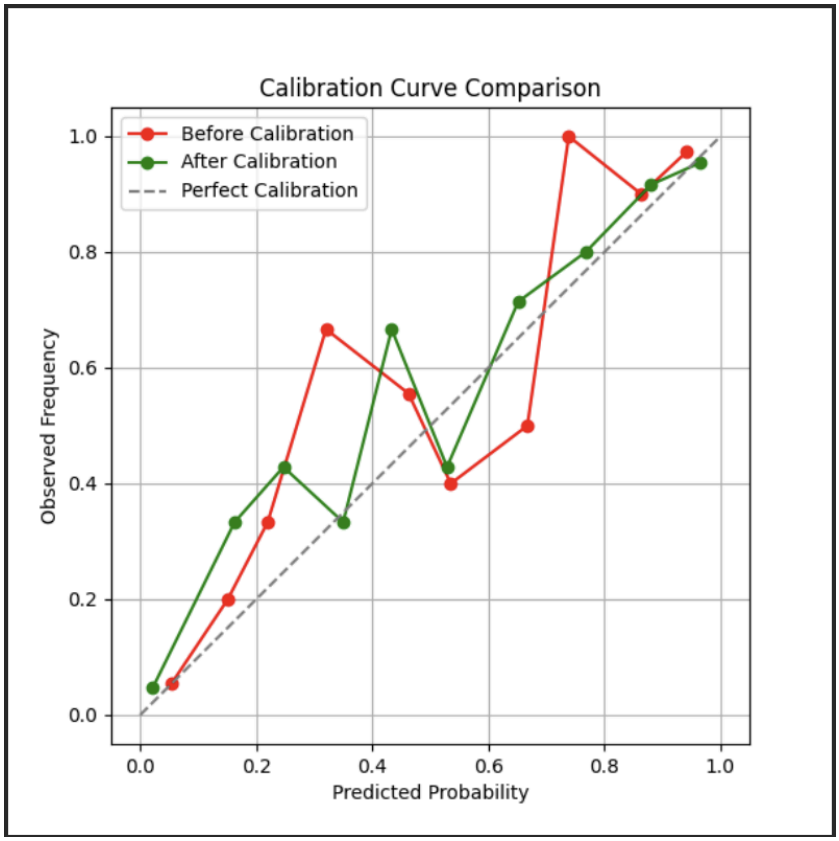
6. CALIBRATION

To improve the reliability of predicted probabilities from each model, a calibration step was conducted using Scikit-learn's CalibratedClassifierCV. The goal of calibration is to ensure that predicted probabilities align closely with actual event frequencies—for example, a predicted 70% risk of heart disease should correspond to a 70% incidence rate among similar patients.

Calibration quality was assessed using the Brier Score, which measures the mean squared error between predicted probabilities and actual outcomes. A lower Brier Score indicates better-calibrated predictions. Depending on the model, either isotonic regression (a non-parametric mapping) or sigmoid (a parametric logistic mapping) was used, along with Stratified K-Fold cross-validation. The calibrated models were then compared to their uncalibrated counterparts to determine whether calibration improved performance and lowered the Brier score, as shown in the GBC results example below.

Kevin O'Donnell
June 1st, 2025

— GBC Model Calibration Results —



Metric	Original	Calibrated
Brier Score	0.1035	0.1015
ROC AUC	0.9311	0.9351

In the case of the Gradient Boosting Classifier (GBC), the calibrated model demonstrated improved calibration and slightly better overall performance, and thus it was selected for downstream use. Calibration consistently improved reliability across all models and was used in all final versions, with the final Calibration strategies:

- XGBoost = isotonic
- GBC = sigmoid
- Logistic Regression = isotonic
- KNN = isotonic

Kevin O'Donnell
June 1st, 2025

7. THRESHOLD TUNING

After calibration, each model underwent additional refinement through both global threshold tuning and subgroup-specific thresholding. The optimal global classification threshold was selected by evaluating F1 scores across a grid of values ranging from 0.20 to 0.75. By default, a threshold of 0.50 means that patients with a predicted probability above 50% are classified as having cardiovascular disease. However, adjusting this threshold allows the model to behave more liberally (lower threshold, more positives identified) or more conservatively (higher threshold, fewer false positives).

To further improve fairness, subgroup-specific thresholds were introduced, particularly for underperforming or underrepresented groups. For these subgroups, a lower threshold than the global baseline was used to increase sensitivity and reduce the chance of false negatives (at the cost of more false positives). For example, if the global threshold is set to 0.50 but the subgroup threshold for female patients is 0.25, then any case identified as female will be evaluated against the 0.25 threshold instead. This is particularly important given that cardiovascular disease is more prevalent in male patients and may be under-predicted in females. By lowering the threshold for the female subgroup, the model becomes more inclusive and equitable in its predictions, improving diversity fairness without compromising overall performance. Each subgroup threshold was also selected using F1 score optimization across a grid search.

8. INDIVIDUAL MODEL TEST SET RESULTS

Each final model was then evaluated on the unseen test set using its calibrated probabilities, optimized global threshold, and any applicable subgroup thresholds. Logistic Regression and XGBoost are the best-performing models exhibiting high F1 scores, and can be used to predict cardiovascular disease in new, unseen individual patients. They can also be further improved with a larger dataset or further optimization rebalancing. The results are summarized in the table below.

Model	Global Threshold	Subgroup Thresholds	ROC AUC	F1 Score	Accuracy	Brier Score	Confusion Matrix (TN/FN /FP/TP)
Logistic Regression	0.6	'Sex_group', 'F': 0.25	0.9517	0.898	0.8897	0.0824	55/6 9/66
KNN	0.45	'Exercise Angina _group', 'Y': 0.65	0.9293	0.8591	0.8456	0.0977	51/10 11/64
GBC	0.35	'Sex_group', 'F': 0.25	0.9351	0.8645	0.8456	0.1013	48/13 8/67
XGBoost	0.45		0.9411	0.9007	0.8897	0.0955	53/8 7/68

Kevin O'Donnell
June 1st, 2025

9. FINAL STACKED MODEL

An ensemble model was also explored by stacking all previously optimized models into a single meta-learner. A stacking ensemble works by learning where each base model performs well or poorly, and then combining their predictions in a way that leverages their individual strengths. For example, if the K-Nearest Neighbors model is particularly effective at predicting outcomes for men over age 65, the stacked ensemble can prioritize its predictions for those specific cases, thus yielding a more accurate overall result. The StackingClassifier was constructed using the final versions of the four base models from earlier in the project: LogisticRegression, K-NearestNeighbors, GradientBoostingClassifier, and XGBoostClassifier. The final estimator (meta-model) was a Logistic Regression classifier tuned using Optuna to maximize F1 score on the validation set. There may be some data leakage or overfitting on the validation set, since the base models were already optimized and exposed to the validation set, however, the dataset was not large enough to hold out an additional validation set just for this step.

A custom auditing script was used throughout the project for overall and subgroup performance evaluation to drive decisions on refinement, weighting, calibration, and threshold tuning. It was applied again to this final stacked model on the unseen test set to assess the final fairness and generalization results. Results of this audit helped verify whether the ensemble provided not only strong overall performance but also equitable treatment across diverse patient groups. For any flagged subgroups, it provides the respective confusion matrix, so the analyst can decipher the root cause of the issue for future refinement.

```
--- Stacked Model Results on Final Test Set ---
Best Params: {'C': 0.645, 'max_iter': 1000}
chosen_global_thresh = 0.60
ROC AUC: 0.9517
F1 Score: 0.8980
Confusion Matrix:
 [[55  6]
 [ 9 66]]
Brier Score: 0.08243552550795459

--- Sex_group ---
Value   N      AUC       F1    Brier Flags
    F  24 0.947368 0.800000 0.069288    OK
    M 112 0.944728 0.905109 0.085253    OK
```

Kevin O'Donnell
June 1st, 2025

--- ExerciseAngina_group ---
Value N   AUC   F1   Brier Flags
    N 87 0.928571 0.779661 0.102751   OK
    Y 49 0.915909 0.977273 0.046364   OK

--- FastingBS_group ---
 Value   N   AUC   F1   Brier Flags
     0 104 0.947173 0.860215 0.087187   OK
     1 32 0.944444 0.962963 0.066994   OK

--- AgeGroup ---
Value N   AUC   F1   Brier Flags
  30s 16 1.000000 0.800000 0.055271   OK
  60s 25 0.855263 0.900000 0.115575   OK
  40s 33 0.982143 0.909091 0.054568   OK
  50s 58 0.942728 0.916667 0.086029   OK

--- RestingBPGroup ---
     Value N   AUC   F1   Brier   Flags
[160, 180) 14 0.866667 0.842105 0.152163 High Brier
[120, 140) 65 0.963542 0.885246 0.075742   OK
[100, 120) 17 0.992857 0.888889 0.090414   OK
[140, 160) 36 0.977273 0.933333 0.070802   OK

--- CholesterolGroup ---
     Value N   AUC   F1   Brier Flags
[250, 300) 24 0.923077 0.846154 0.098621   OK
[200, 250) 68 0.953140 0.906977 0.088102   OK
[150, 200) 23 0.960784 0.909091 0.053924   OK
[300, 350) 11 1.000000 1.000000 0.025579   OK

--- MaxHRGroup ---
     Value N   AUC   F1   Brier Flags
[150, 180) 48 0.945055 0.666667 0.093102 Low F1
 [90, 120) 21 0.844444 0.903226 0.105138   OK
[120, 150) 60 0.943182 0.943820 0.072251   OK
  --- Confusion Matrix for MaxHRGroup: [150, 180) (Flagged) ---
  True Negative (0,0): 34, False Positive (0,1): 1
  False Negative (1,0): 6, True Positive (1,1): 7
[[34 1]
 [6 7]]

Kevin O'Donnell
June 1st, 2025

--- OldpeakGroup ---
Value N AUC F1 Brier Flags
[0.0, 1.0) 78 0.962847 0.857143 0.075850 OK
[1.0, 2.0) 32 0.883117 0.857143 0.132959 OK
[2.0, 3.0) 15 1.000000 1.000000 0.017870 OK

--- ChestPainType_group ---
Value N AUC F1 Brier Flags
NAP 32 0.931250 0.833333 0.094838 OK
ASY 71 0.914456 0.920354 0.085293 OK
ATA 28 1.000000 1.000000 0.016930 OK

--- RestingECG_group ---
Value N AUC F1 Brier Flags
LVH 18 0.883117 0.705882 0.178807 High Brier
Normal 91 0.960049 0.907216 0.080430 OK
ST 27 1.000000 0.969697 0.024948 OK

--- ST_Slope_group ---
Value N AUC F1 Brier Flags
Up 63 0.945385 0.600000 0.079618 Low F1
Flat 61 0.852594 0.944444 0.085158 OK
Down 12 0.925926 0.947368 0.083392 OK
--- Confusion Matrix for ST_Slope_group: Up (Flagged) ---
True Negative (0,0): 49, False Positive (0,1): 1
False Negative (1,0): 7, True Positive (1,1): 6
[[49 1]
[7 6]]

--- Overall Model Performance ---
Confusion matrix:
[[55 6]
[9 66]]
Classification report:

	precision	recall	f1-score	support
0	0.86	0.90	0.88	61
1	0.92	0.88	0.90	75
accuracy			0.89	136
macro avg	0.89	0.89	0.89	136
weighted avg	0.89	0.89	0.89	136

Kevin O'Donnell
June 1st, 2025

Accuracy: 0.8897
ROC AUC (probabilities): 0.9517
Brier Score (probabilities): 0.0824

--- Error Analysis of Misclassified Samples ---
Total True Negatives (TN): 55
Total False Positives (FP): 6
Total False Negatives (FN): 9
Total True Positives (TP): 66

--- Descriptive Statistics for False Positives (FP) ---

	Age	Sex	...	y_proba	y_pred_subgrp
count	6.000000	6.000000	...	6.000000	6.0
mean	62.833333	0.833333	...	0.805012	1.0
std	4.996666	0.408248	...	0.092686	0.0
min	58.000000	0.000000	...	0.675951	1.0
25%	59.750000	1.000000	...	0.765758	1.0
50%	62.000000	1.000000	...	0.790557	1.0
75%	63.500000	1.000000	...	0.848920	1.0
max	72.000000	1.000000	...	0.947143	1.0

--- Descriptive Statistics for False Negatives (FN) ---

	Age	Sex	...	y_proba	y_pred_subgrp
count	9.000000	9.000000	...	9.000000	9.0
mean	49.222222	0.888889	...	0.376727	0.0
std	10.556725	0.333333	...	0.150967	0.0
min	34.000000	0.000000	...	0.111265	0.0
25%	43.000000	1.000000	...	0.245103	0.0
50%	50.000000	1.000000	...	0.436618	0.0
75%	55.000000	1.000000	...	0.436618	0.0
max	66.000000	1.000000	...	0.577599	0.0

[8 rows x 22 columns]

--- SHAP Summary Plot for Overall Feature Importance ---

Top 10 features by importance:
ST_Slope_Up 0.472203
ChestPainType_ASY 0.471591
ST_Slope_Flat 0.411336
Oldpeak 0.398454
ExerciseAngina 0.376297

```
Sex                 0.344455
FastingBS           0.285880
MaxHR               0.190866
ChestPainType_ATA   0.168987
ChestPainType_NAP   0.142199
dtype: float64

Bottom 10 features by importance:
ChestPainType_ATA   0.168987
ChestPainType_NAP   0.142199
Cholesterol         0.137062
Age                 0.124388
RestingBP           0.094273
RestingECG_ST       0.022359
ST_Slope_Down       0.019578
RestingECG_LVH      0.014191
ChestPainType_TA    0.003710
RestingECG_Normal   0.003034
dtype: float64

Highly correlated pairs (>|0.85|):
  None
```