Cisse, Yelene

Dr. Spencer W. Luo

COMS 4995W32

7 October 2025

Heart Disease Prediction: Comparative Analysis of Classification Models

Introduction

This report presents a comprehensive analysis of machine learning models applied to heart disease prediction using the UCI Heart Disease dataset. The primary objective is to compare the performance of three distinct classification approaches: Bernoulli Naive Bayes (BernoulliNB), Gaussian Naive Bayes (GaussianNB), and Linear Regression with various regularization techniques.

The dataset, sourced from the UCI Machine Learning Repository, contains medical information about 303 patients with 14 attributes including both continuous clinical measurements (age, resting blood pressure, cholesterol, maximum heart rate, ST depression) and categorical health indicators (sex, chest pain type, fasting blood sugar, exercise-induced angina). The target variable indicates the presence of heart disease, defined as greater than 50% diameter narrowing in any major vessel (Detrano et al.).

Our analysis evaluates how each model responds to different feature sets and preprocessing approaches, with the goal of identifying the most effective model for predicting heart disease presence in patients. The models examined in this report are:

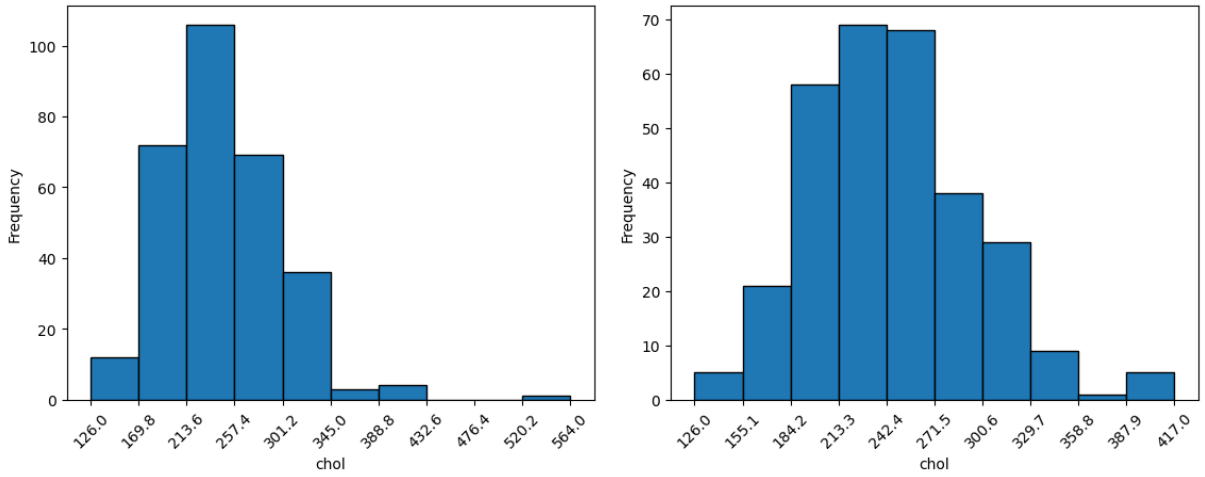
- **BernoulliNB**: A probabilistic classifier designed for binary/categorical features, tested with varying Laplace smoothing parameters (alpha = 0.00001, 0.01, 0.05, 1)
- **GaussianNB**: A probabilistic classifier that assumes continuous features follow a Gaussian distribution
- **Linear Regression**: Implemented with three regularization approaches (regular, Ridge with alpha=1, and LASSO with alpha=0) to assess their impact on prediction performance and feature importance

1. Data Exploration

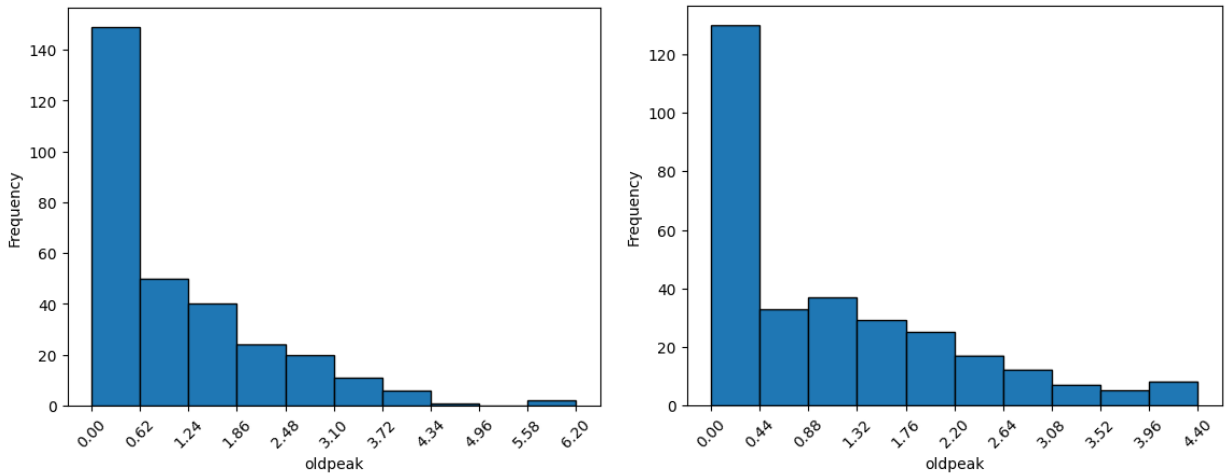
Initial data exploration began with examining hit rates and distributions for all variables using summary statistics and histogram visualizations. The hit rate analysis revealed that only two variables had missing values: 'ca' (number of major vessels, 98.7% complete) and 'thal' (thalassemia, 99.3% complete), with all other variables having 100% complete data.

Histogram analysis of continuous variables identified outliers requiring attention. One patient had a cholesterol value of 564 mg/dl (well above the 99th percentile of 406.7), and two patients

had oldpeak values exceeding 5.0 (5.6 and 6.2 compared to 99th percentile of 4.2). Capping was applied to these outliers using the maximum non-outlier values to prevent them from disproportionately influencing model predictions later. Below is a side by side comparison of distributions with and without capping:



Before vs After capping Cholesterol



Before vs After capping Oldpeak

For missing values, 'ca' was imputed with the median value of 0 (which was also the mode), while 'thal', being categorical, was filled with -1 to create a distinct "unknown" category that could be one-hot encoded without losing information.

Correlation analysis of continuous variables showed that features were not closely correlated with each other (all correlations < 0.4 except age-thalach at -0.394), allowing all variables to be included in models without concern for competing signals. The variables most strongly correlated with the target were: ca (0.519), oldpeak (0.504), and thalach (-0.415).

	age	trestbps	chol	thalach	oldpeak	ca	num
age	1	0.284946	0.20895	-0.39381	0.203805	0.362605	0.222853
trestbps	0.284946	1	0.13012	-0.04535	0.189171	0.098773	0.157754
chol	0.20895	0.13012	1	-0.00343	0.046564	0.119	0.070909
thalach	-0.39381	-0.04535	-0.00343	1	-0.34309	-0.26425	-0.41504
oldpeak	0.203805	0.189171	0.046564	-0.34309	1	0.295832	0.504092
ca	0.362605	0.098773	0.119	-0.26425	0.295832	1	0.518909
num	0.222853	0.157754	0.070909	-0.41504	0.504092	0.518909	1

Correlation Matrix of Continuous Variables

2. Feature Engineering

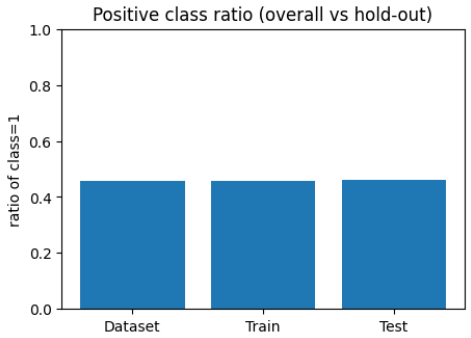
Feature engineering focused on transforming continuous variables into categorical representations suitable for BernoulliNB:

- Blood pressure was categorized as normal (≤130 mm Hg) or high (>130 mm Hg) based on CDC guidelines.
- Age was grouped into life stages: young adulthood (<40), middle adulthood (40-59), young old age (60-74), and old old age (≥75).
- Cholesterol was classified using John Hopkins guidelines as normal (<200 mg/dL), borderline high (200-239 mg/dL), or high (≥240 mg/dL).
- For oldpeak, quantile-based binning starting at the 50th percentile (since 50% of values were 0) created 6 categories, while thalach was divided into deciles to capture heart rate ranges.

For data transformations before model training, one-hot encoding was applied to all categorical variables to create binary indicator features, and standardization (zero mean, unit variance) was applied to continuous variables for GaussianNB and Linear Regression models.

3. Model Training

The dataset was split 70-30 into training (212 observations) and testing (91 observations) sets using stratification with random_state=5. Stratification maintained consistent proportions of heart disease positive cases: 45.9% overall, 45.8% training, and 46.2% testing.



Each model was trained twice: first with features naturally suited to the model type, and second with all features after applying appropriate transformations. For BernoulliNB, the first run used only raw categorical variables, while the second included all feature-engineered categorical derivations. For GaussianNB and Linear Regression, the first run used only standardized continuous variables, while the second included continuous plus one-hot encoded categorical features.

Different hyperparameters were tested for each model type. BernoulliNB was evaluated with four Laplace smoothing values (alpha = 0.00001, 0.01, 0.05, 1). Linear Regression was tested in three variants: regular, Ridge (alpha=1), and LASSO (alpha=0).

4. Model Evaluation

Performance by Model Type

BernoulliNB performed significantly better with all feature-engineered variables (accuracy 0.8352) compared to raw categorical variables only (accuracy ranging from 0.7912 to 0.8022 depending on alpha). This improvement was supported by increased true predictions in the confusion matrix (76/91 vs 72-73/91) and consistent ROC/AUC of 0.8350. Interestingly, once all features were included, the model became completely insensitive to the Laplace smoothing parameter—all alpha values produced identical results.

GaussianNB also showed dramatic improvement with all features (accuracy 0.7692) versus continuous variables only (accuracy 0.4505), demonstrating the critical importance of categorical information for heart disease prediction. The confusion matrix improved from approximately 41/91 to 69/91 correct predictions, and AUC increased substantially to 0.7670.

Linear Regression variants showed minimal difference between continuous-only and all-features configurations, with both achieving same accuracy. This suggests that standardized continuous variables already captured most of the predictive signal for linear models.

Comparing the Naive Bayes models, BernoulliNB (0.8352 accuracy, 0.8350 AUC) outperformed GaussianNB (0.7692 accuracy, 0.7670 AUC) by 6.6 percentage points. This indicates that the Bernoulli distribution's assumptions for binary/categorical data better match this medical dataset's characteristics than GaussianNB's assumption of normally distributed continuous features.

model	alpha	accuracy_run1	accuracy_run2	% Diff	roc/auc_run1	roc/auc_run2	%Diff
BernoulliNB	1	0.8022	0.8352	4%	0.8027	0.835	4%
BernoulliNB	0.01	0.7912	0.8352	5%	0.7925	0.835	5%
BernoulliNB	0.05	0.8022	0.8352	4%	0.8027	0.835	4%
BernoulliNB	0.00001	0.7912	0.8352	5%	0.7925	0.835	5%
GaussianNB	n/a	0.4505	0.7692	41%	0.4881	0.767	36%
LinearReg	n/a	0.8242	0.8242	0%	0.9082	0.9033	-1%
LinearReg	0	0.5385	0.5385	0%	0.5	0.5	0%
LinearReg	1	0.8242	0.8242	0%	0.9106	0.9096	0%

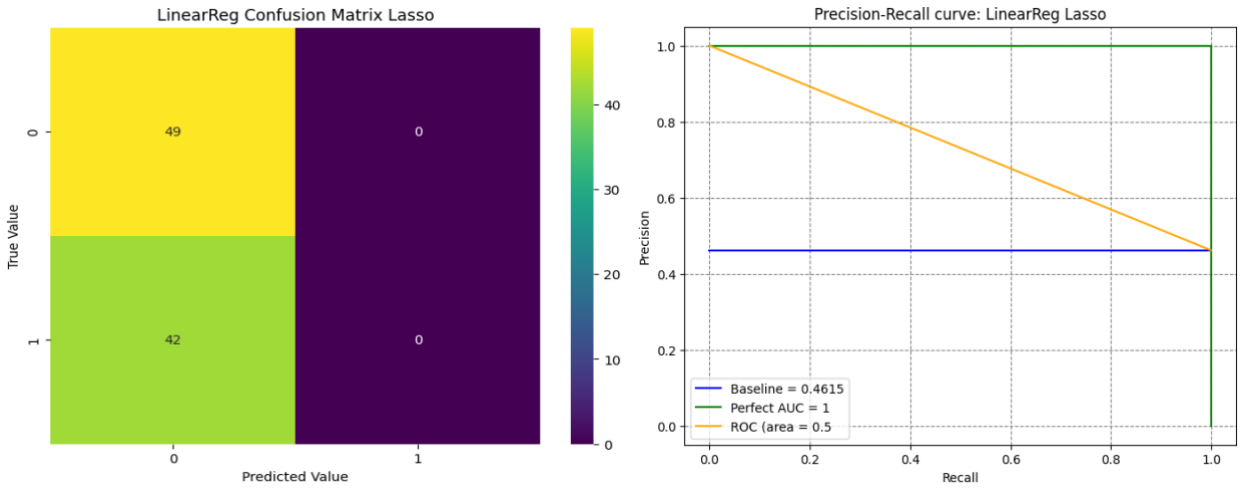
Linear Regression Regularization Comparison

Among the three Linear Regression variants, **Ridge** provided the best overall results across all metrics. Ridge achieved accuracy of 0.8242, the highest AUC of 0.9096, and optimal error metrics (MSE: 0.124, MAE: 0.276, R^2: 0.501). Regular Linear Regression performed similarly in accuracy (0.8242) but had slightly higher error (MSE: 0.129, MAE: 0.282, R^2: 0.480) and lower AUC (0.9033).

model	MSE	MAE	R2
LinearReg_Regular LR	0.129350521	0.282304019	0.479518144
LinearReg_Lasso	0.24853664	0.496734398	-6.41009E-05
LinearReg_Ridge	0.124122033	0.27563169	0.500556583

Metrics Table Summary for Linear Regression

LASSO produced the worst results, with accuracy of only 0.5385, negative R^2 (-0.000064), and AUC of 0.5 (random guessing). The confusion matrix revealed LASSO predicted all 91 test cases as the negative class, completely failing to identify any diseased patients.

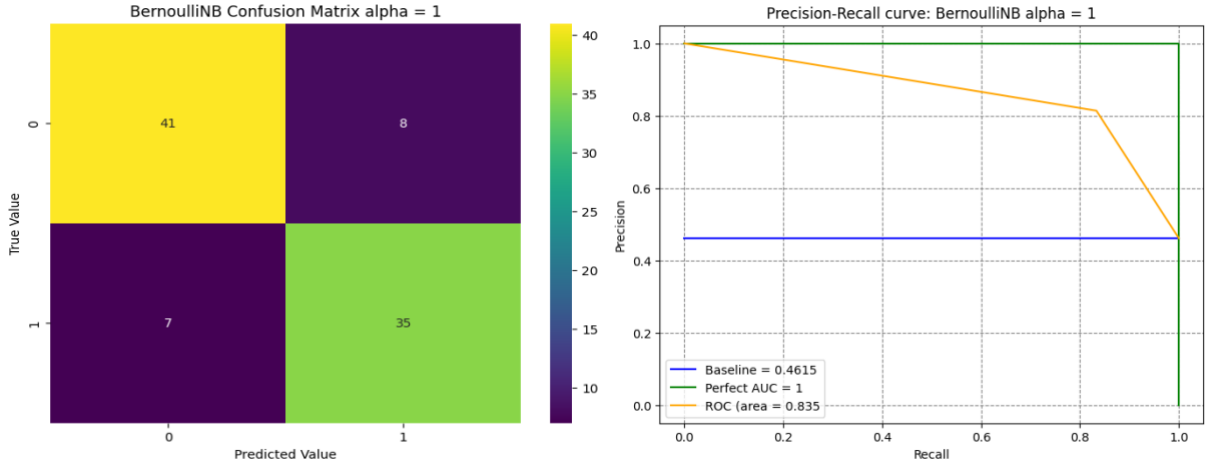


Examination of the coefficient table showed LASSO shrunk all variables to zero or near-zero, eliminating the predictive signal entirely. This was surprising given LASSO's theoretical strength in feature selection—some coefficients should have remained non-zero to identify important predictors. The extreme over-regularization, despite alpha=0, rendered LASSO unsuitable for this application.

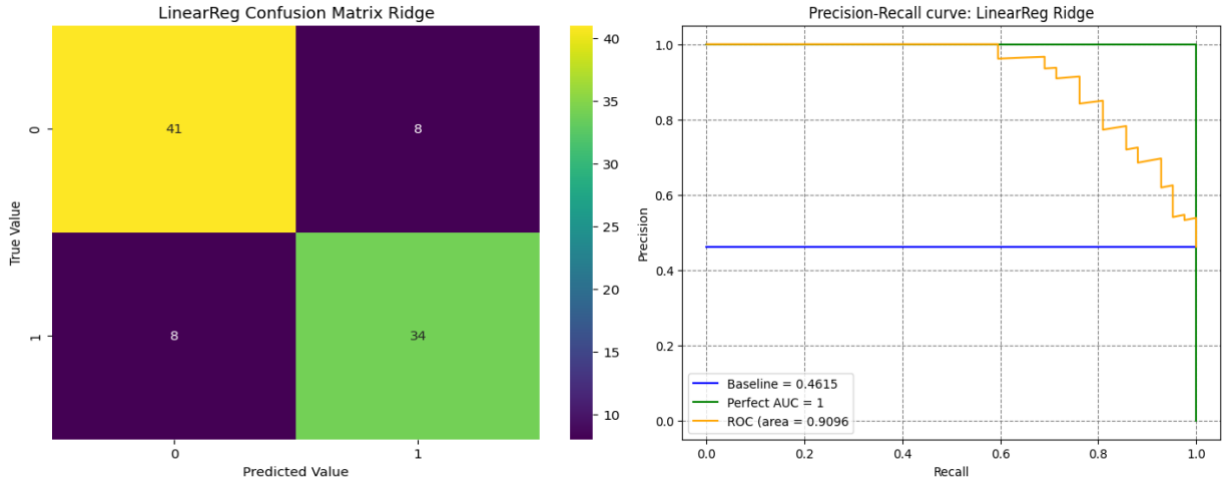
Best Model Selection

For fair comparison, we evaluated the best configuration from each model type: BernoulliNB with alpha=1 and all features, GaussianNB with all features, and Linear Regression Ridge with all features.

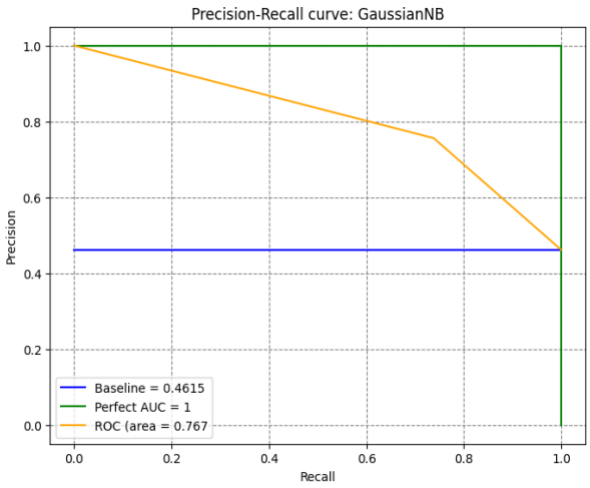
BernoulliNB emerged as the best overall model for this exercise with 0.8352 accuracy (76/91 correct). The confusion matrix showed 41 true positives, 35 true negatives, 8 false positives, and 7 false negatives, representing a balanced 14.3% miss rate and 16.3% false alarm rate suitable for clinical screening.

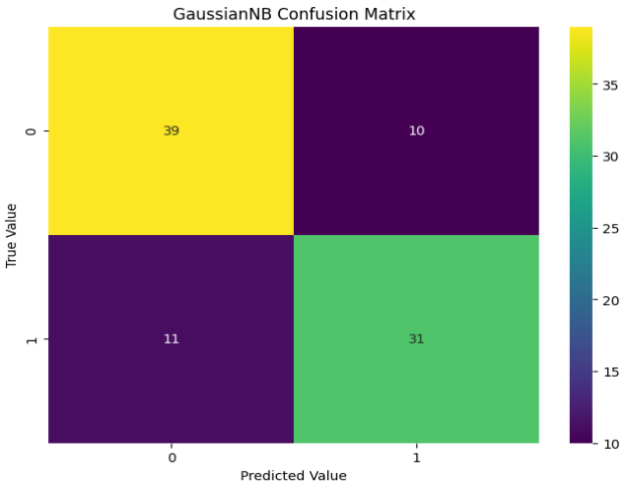


Ridge Regression placed second with 0.8242 accuracy (75/91 correct) but demonstrated superior discriminative ability with the highest AUC of 0.9096—a 7.5 percentage point advantage over BernoulliNB (0.8350). This indicates Ridge produces better-calibrated probability estimates, making it valuable when rank-ordering patients by risk or adjusting decision thresholds for different clinical contexts.



GaussianNB ranked third with 0.7692 accuracy (69/91 correct) and AUC of 0.7670, showing it is serviceable but outperformed by both BernoulliNB and Ridge.





Weighted average metrics confirmed these rankings. BernoulliNB achieved precision 0.836, recall 0.835, and F1 0.835. Ridge had precision 0.824, recall 0.824, and F1 0.824. GaussianNB showed precision 0.769, recall 0.767, and F1 0.767.

model_type	metrics	weighted avg
BernoulliNB_alpha = 1	precision	0.835606738
BernoulliNB_alpha = 1	recall	0.835164835
BernoulliNB_alpha = 1	f1-score	0.835284788
LinearReg_Ridge	precision	0.824175824
LinearReg_Ridge	recall	0.824175824
LinearReg_Ridge	f1-score	0.824175824
GaussianNB	precision	0.768968105
GaussianNB	recall	0.769230769
GaussianNB	f1-score	0.769006094

6. Conclusion

This analysis demonstrated that thoughtful feature engineering dramatically improves model performance, particularly for algorithms with specific data type requirements. BernoulliNB benefited most from categorical derivations of continuous variables, achieving the highest accuracy of 83.5% and proving robust across all Laplace smoothing values once properly engineered features were provided. The 41-feature engineered dataset provided sufficient information density that regularization parameter variations became inconsequential.

Ridge Regression, while slightly behind in accuracy, still did well in probability calibration with AUC of 0.9096. It would be preferred over GaussianNB for this exercise given that it surpassed the model in each category. The gentle L2 regularization improved generalization while preserving all feature information.

LASSO's failure highlights the importance of careful hyperparameter selection and the potential dangers of extreme regularization, even with theoretically minimal alpha settings. Future work could explore cross-validated LASSO implementations to determine if appropriate regularization strength exists for this problem.

For clinical deployment, BernoulliNB with feature-engineered categorical variables (alpha=1) is recommended for maximum classification accuracy in screening contexts.

AI Tool Usage Disclosure

This analysis involved the use of several AI tools throughout the workflow, with transparency about their contributions and limitations:

ChatGPT was utilized for:

- Feature engineering recommendations, particularly suggestions for medically meaningful binning strategies for continuous variables (blood pressure categories, cholesterol levels)
- Code snippets for data preprocessing, model training, and evaluation that were subsequently edited for accuracy and adapted to the specific dataset requirements and use case
- Understanding conceptual differences between ROC and AUC metrics, specifically clarifying baseline calculations for each curve type
- Debugging error messages during model execution (missing packages identified, parentheses missing, figure saving errors).

Code Adaptation Requirements: AI-generated code required manual adjustments due to package incompatibilities. For example, when exporting results to Excel, ChatGPT initially provided code snippets for xlsxwriter, but when prompted for methods to add new worksheets with plots, it provided syntax for openpyxl instead. This mismatch required manual debugging and rewriting to maintain consistency with the xlsxwriter package used throughout the analysis.

Report Creation Process: The final report was drafted using a two-step AI-assisted process:

1. Initial observations and model comparisons were captured using Microsoft Word's dictate feature to transcribe spoken analysis of results viewed in Excel outputs
2. The transcript was then passed to Claude, along with the PDF of annotated code, to structure findings into a formal report following a detailed self-written outline
3. The report was manually edited to integrate relevant plots, tables, correct information summarization, and add specific references with actual values from the results

Human Contributions: All feature engineering decisions, model comparisons, and interpretation of results were made independently after reviewing the data characteristics and medical literature. The choice of which categorical bins to create (e.g., quantile-based for oldpeak due to its 50% zero values, clinical guidelines for blood pressure and cholesterol) came

from domain research rather than AI suggestions. AI tools served as accelerators for code implementation and documentation structure, but core analytical insights, model selection criteria, and the interpretation of results based on pipeline output were self-derived.

Works Cited (Includes code references)

Detrano, R., et al. "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease." *American Journal of Cardiology*, vol. 64, 1989, pp. 304-310.

"Heart Disease." *UCI Machine Learning Repository*, 2023, archive.ics.uci.edu/dataset/45/heart+disease. Accessed 6 Oct. 2025.

“High Blood Pressure.” Centers for Disease Control and Prevention, 15 Jan. 2025, https://www.cdc.gov/high-blood-pressure/index.html. Accessed 6 Oct. 2025.

Hurley, Margaret Anne. “A Reference Relative Time-Scale as an Alternative to Chronological Age for Cohorts with Long Follow-Up.” Emerging Themes in Epidemiology, vol. 12, 2015, doi:10.1186/s12982-015-0043-6. PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC4684933/. Accessed 6 Oct. 2025.

“Chronological Age.” ScienceDirect Topics, Elsevier, https://www.sciencedirect.com/topics/computer-science/chronological-age. Accessed 6 Oct. 2025.

“Lipid Panel.” Johns Hopkins Medicine, Johns Hopkins University, https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel#:~:text=Normal:%20Less%20than%20150%20mg,14%20hours%20before%20this%20test. Accessed 6 Oct. 2025.