Predicting Risk of Heart Disease for Early Detection: A Machine-Learning Approach

Larry Vue, DS785

Fall Semester 2025

Abstract:

Cardiovascular disease (heart disease) is still the leading cause of death worldwide and significantly impacts healthcare expenses. With a better understanding of data and newer developments in data science, the need for early detection can save lives and reduce healthcare costs [1][2]. This project developed and evaluated supervised learning models to predict heart disease using clinical variables. The data primarily come from the UCI Cleveland Heart Disease dataset and a larger Kaggle cardiovascular dataset. Data cleaning, exploratory analysis, and feature engineering were used to assess the predictive value of key risk factors. The data helped compare baseline and advanced machine learning classifiers, including logistic regression, decision trees, random forests, and gradient boosting (XGBoost), using stratified cross-validation. The focus, aligned with business metrics, is to prioritize recall to minimize false negatives. Class imbalance was addressed using class weights and thresholds, along with ROC/PR analysis and cost-sensitive decision-making. Results show that the interpretations of tree-based models align with clinically relevant relationships. The final model will be a calibrated logistic regression. This achieved strong ranking performance on the internal test set and an interpretable coefficient profile for clinicians. Error analysis revealed that false negatives are often in hard-to-see cases; lowering the threshold slightly reduced misses while maintaining acceptable precision. Overall, the project demonstrates a simple, interpretable model that can provide actionable risk. The model will show the importance of threshold choices for clinical workflows.

1. Problem Definition:

Cardiovascular disease (CVD) caused an estimated 19.8 million deaths in 2022, accounting for approximately 32% of all global deaths [1]. In the United States, CVD was responsible for more than 938,000 deaths in 2020 and remains the leading cause of mortality. The economic impact is substantial, with direct and indirect costs exceeding 400 billion dollars [2]. Leveraging machine learning to early-detect high-risk patients using clinical data has the potential to significantly benefit health systems.

This project addresses the following business problem: how can clinics or healthcare systems utilize routine clinical variables to identify patients at elevated risk of heart disease? Primary care and cardiology clinics routinely collect standard risk factors for coronary artery disease, including age, blood pressure, cholesterol, exercise response, and symptoms. However, risk assessments are frequently informal or rely on paper-based tools. Many clinics and hospitals lack systematic methods to integrate these factors into consistent, individualized risk scores that can be tracked over time and used for patient outreach.

The objective of this project is to develop a binary classification model that predicts the presence of clinically significant heart disease using tabular variables commonly available in clinical workflows. The model is designed to output calibrated disease probabilities that can be translated into one or more decision thresholds. The primary aim is to minimize missed true cases of heart disease, as these are more costly than unnecessary testing for low-risk patients. Therefore, the model prioritizes high recall

(sensitivity) while maintaining acceptable precision. A secondary objective is to ensure model interpretability, enabling clinicians and managers to understand the major drivers of heart disease risk and to facilitate adoption in practice.

The primary users of this model are clinicians, care management teams, and operational leaders. Clinicians may incorporate risk scores as one factor in decisions regarding diagnostic tests or imaging. Population health and care management teams can apply the model to identify high-risk individuals for outreach or counseling. Operational leaders may analyze the distribution of predicted risk and assess the impact of alternative thresholds to inform capacity planning and cost prediction. The scope of this project is limited to open datasets, and any real-world deployment would necessitate a separate governance and ethical review process.

Success was defined using both statistical and practical criteria. The target was to achieve an ROC-AUC of at least 0.85 on the test set, with recall of at least 0.85 at a threshold where precision exceeds 0.75. From a calibration perspective, predicted probabilities were expected to align closely with observed event rates across deciles of risk. The model must also be straightforward to implement in a production environment, with easily interpretable thresholds.

2. Data Collection, Cleaning, and Preprocessing

The primary dataset that will be studied is the Cleveland heart disease subset of the UCI Machine Learning Repository [3]. It contains 303 patients evaluated for having an underlying risk of coronary artery disease. They are either diagnosed late or remain undiagnosed until they present with acute events. Many predictor variables are included, such as age, sex, chest pain type, resting systolic blood pressure, serum cholesterol, fasting blood sugar, resting ECG interpretation, maximum heart rate during exercise, exercise-induced angina, ST depression (oldpeak), slope of the ST segment, number of major vessels, and thalassemia status [3]. The original outcome num ranges from 0 (no disease) to 4 (severe disease); for this project, it was converted to a binary target that equals 1 whenever num > 0, representing any angiographically defined disease, consistent with prior classification studies using this dataset [7], [8].

The raw file uses “?” to denote missing values, particularly in ‘ca’ and ‘thal’ [3]. Data were loaded with pandas, and the missing values were specified. Because the fraction of missing entries is small, and following common practice in prior work with this dataset [3], [7], missing values were imputed with the mode of each affected variable. After imputation, the original ‘num’ field was dropped. The resulting class balance was approximately 54% without disease and 46% with disease. Descriptive statistics for several key predictors are shown in Table 1.

Table 1. Descriptive stats for selected Cleveland Predictors (n=303)

Variable	Mean	Std Dev	Min	25%	50%	75%	Max
age	54.44	9.04	29	48	56	61	77
sex	0.68	0.47	0	0	1	1	1
cp	3.16	0.96	1	3	3	4	4
trestbps	131.69	17.60	94	120	130	140	200
chol	246.69	51.78	126	211	241	275	564

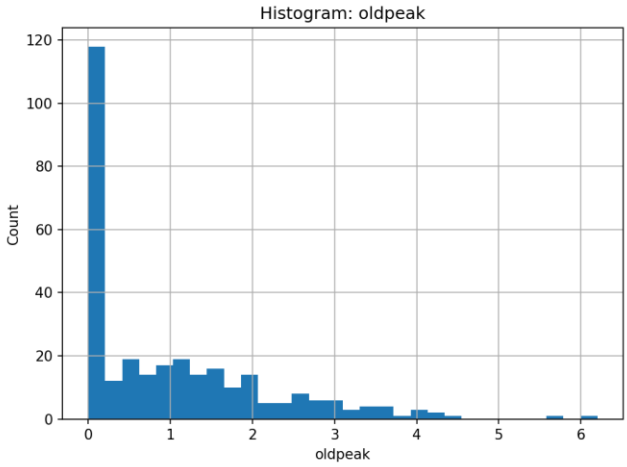
Within the Cleveland dataset, the emphasis was on handling missing values rather than trimming outliers, because the sample size is modest and most values already fall in plausible clinical ranges. For modeling, a consistent preprocessing pipeline was essential. The full Cleveland dataset was split into training and test subsets using a stratified 75-25 split. These sets show that the proportion of patients without disease is similar between the training and test sets. Exploratory data analysis used variables on their original scales to simplify interpretation. For supervised learning, numerical features such as age, resting systolic blood pressure, cholesterol, maximum heart rate (thalach), ST-segment depression (oldpeak), and several engineered variables were standardized using z-score normalization. In contrast, categorical features were one-hot encoded, dropping one category per feature. These transformations were applied to a scikit-learn 'ColumnTransformer' and combined with a classifier in a single Pipeline. The same preprocessing steps were consistently applied during cross-validation, model fitting, and test-set evaluation [6].

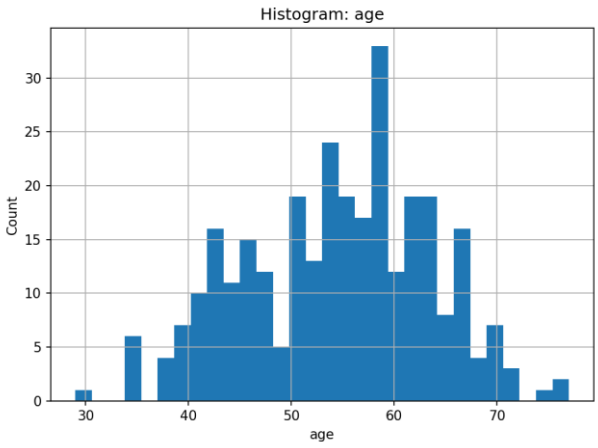
To explore a simple form of external validation, a reduced feature set was extracted from the Cleveland data, consisting of age, sex, and resting systolic blood pressure, as these are common in many clinical datasets. The Kaggle dataset included fields for age in years, a binary gender indicator that could be mapped to the Cleveland sex coding, and systolic blood pressure. Aligning these fields enabled training and reducing the logistic model on the Cleveland data and applying it directly to the Kaggle data without an extensive schema. A separate script handled this reduced-feature external test and produced the internal–external performance comparison and the corresponding calibration plot for the reduced model.

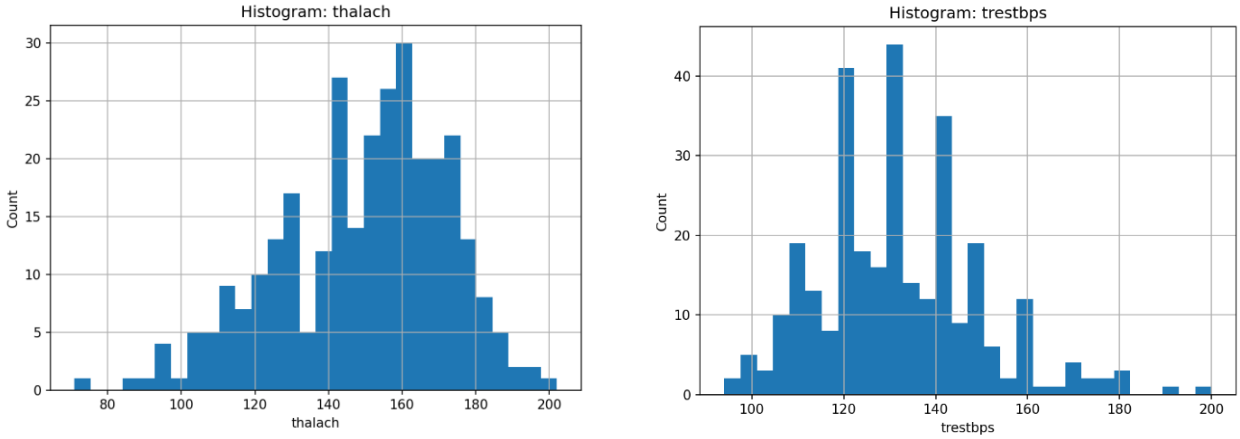
3. Exploratory Data Analysis and Feature Engineering

Exploratory analysis focused on understanding variable distributions and how key predictors relate to the binary target. Histograms for age, resting systolic blood pressure, cholesterol, maximum heart rate, and ST depression showed plausible ranges and modest skew (Figure 1). This is consistent with descriptive statistics reported in earlier analyses of the Cleveland data [3], [7]. Age was centered in the mid-50s, with a tail into older ages, indicating greater heart disease risk. Resting blood pressure and cholesterol were mildly right-skewed. ST depression was low for most patients, but with a small group showing larger changes.

Figure 1:

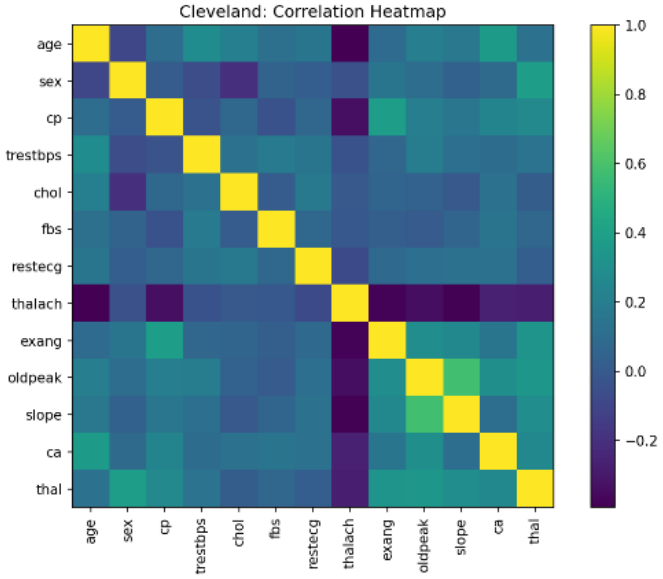






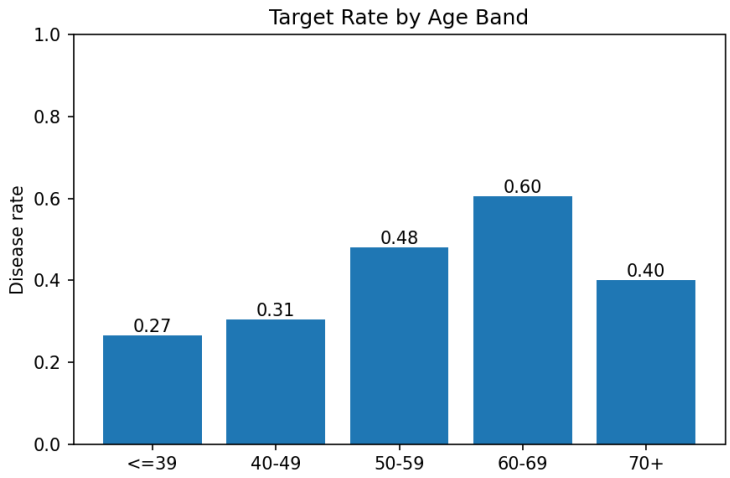
A correlation heatmap of numerical predictors (Figure 2) highlighted relationships among the variables in the Cleveland dataset. Age was modestly correlated with resting blood pressure and cholesterol, suggesting a higher likelihood of cardiovascular risk factors with increasing age [2]. Maximum heart rate was negatively associated with age and positively associated with exercise capacity. ST depression tended to increase with age and resting blood pressure. No pair of predictors exhibited extreme correlation, suggesting that the full feature set could be retained without severe multicollinearity, consistent with prior model-building work on this dataset [7], [8].

Figure 2:



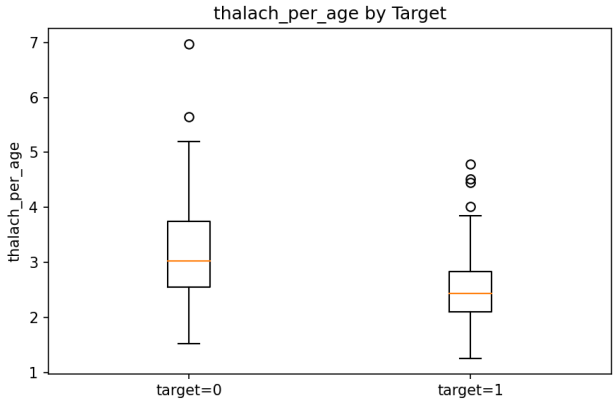
Age was binned into ordered bands (≤39, 40–49, 50–59, 60–69, 70+) to capture non-linear age effects in an interpretable way. The proportion of patients with disease rose steadily across age bands (Figure 3), supporting the use of an age-band feature with a simple monotonic risk pattern and aligning with epidemiologic patterns for coronary disease [1], [2].

Figure 3:



Several engineered features were added. A fitness proxy, thalach_per_age, was defined as maximum heart rate divided by age. Boxplots (Figure 4) show that patients with heart disease tend to have lower values of this ratio, consistent with reduced exercise capacity. Additional features included cholesterol per age (chol_per_age) and the interaction oldpeak_x_exang, which is large when ischemic ST changes and exercise-induced angina co-occur—patterns that are consistent with clinical knowledge of exertional ischemia [2], [3].

Figure 4:



Group differences between patients with and without disease were quantified using Welch's t-tests for key numerical predictors. Maximum heart rate was significantly lower in patients with disease, while ST depression, age, and resting systolic blood pressure were significantly higher (Table 2). These results aligned with clinical expectations and with previous studies that found similar discriminative patterns in the Cleveland dataset [7], [8], motivating their central role in the model.

Table 2. Group comparison for continuous Predictors (Cleveland Dataset)

Feature	**Mean (no disease)**	**Mean (disease)**	**p-value**
thalach	158.38	139.26	1.86e-13
oldpeak	0.59	1.57	6.81e-13
age	52.59	56.63	3.92e-05
trestbps	129.25	134.57	2.60e-02
chol	242.64	251.47	3.54e-02

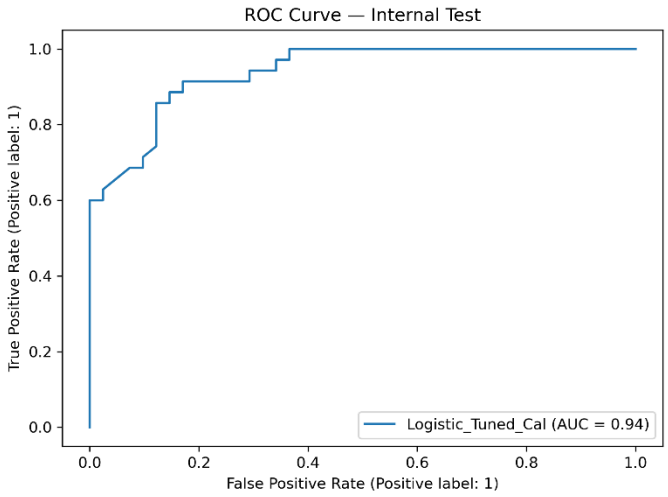
4. Model Development and Evaluation

The final modeling approach used was the logistic regression classifier with a scikit-learn pipeline [6]. Categorical features (chest pain type, rest ECG, slope, number of vessels, thalassemia, and age band) were one-hot encoded, and numerical features (age, sex, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, ST depression, and engineered variables) were standardized. The classifier used ℓ2-regularization with the inverse regularization strength C tuned via grid search over a small set of candidate values, a common strategy for logistic models in medical prediction tasks [7], [8].

In addition to logistic regression, several other supervised learning algorithms were evaluated as secondary baselines, including k-nearest neighbors (KNN), random forest, and gradient boosting / XGBoost. These models were fit using the same train–test split and a similar preprocessing strategy (scaling of numerical features and one-hot encoding of categorical variables). On this relatively small dataset, tree-based ensembles and XGBoost achieved ROC-AUC values that were comparable to, but not substantially better than, the tuned logistic regression model, while KNN was more sensitive to feature scaling and offered no clear advantage in discrimination. Because logistic regression provided competitive performance, produced well-behaved calibrated probabilities after isotonic calibration, and yielded coefficients that can be directly interpreted as risk directions and relative strengths, it was selected as the primary model for this project [5], [6], [8].

Stratified five-fold cross-validation was used for hyperparameter tuning on the training set with ROC-AUC as the primary score. Modest regularization provided the best trade-off between fit and generalization. The primary goal was to emphasize ranking and to catch high-risk patients; recall was treated as more important than precision. The Cleveland test set with a tuned logistic model achieved an ROC-AUC of around 0.94 (Figure 5). This shows a strong ability to separate patients with heart disease from those without heart disease. This is comparable to other studies that have used the Cleveland dataset and have shown similar results [7], [8].

Figure 5.



To support threshold selection, performance was summarized at three candidate probability cutoffs. I used the cutoffs of 0.50, 0.45, and 0.40. Table 3 shows that lowering the threshold from 0.50 to 0.45 increases recall with only a modest loss in precision. Lowering to 0.40 yields diminishing gains in recall. A threshold of 0.45 was therefore selected as the primary operating point. This yielded a recall near 0.89 and precision near 0.84, while maintaining an overall ROC-AUC of roughly 0.94. This choice reflects a recall-first goal that may be appropriate when missed cases of heart disease are of higher importance.

Table 3:

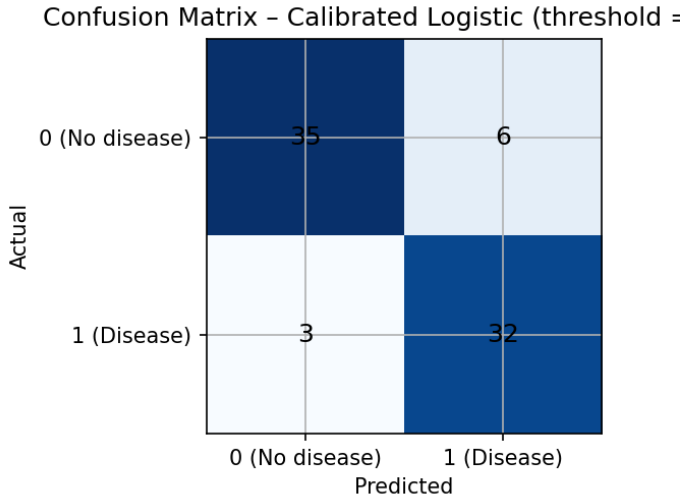
Dataset	Threshold	ROC-AUC	Recall	Precision	F1
Internal (0.50)	0.50	0.935	0.86	0.83	0.85
Internal (0.45)	0.45	0.935	0.89	0.84	0.86
Internal (0.40)	0.40	0.935	0.89	0.84	0.86

To obtain more reliable probability estimates, the tuned pipeline was further calibrated using isotonic regression with cross-validation on the training data [5]. Isotonic calibration is well-suited when the classifier is strong but may be mis-calibrated and has been shown to improve probability estimates across several supervised learning methods [5].

5. Results, Insights, and Recommendations

With the chosen operating threshold of 0.45, the model's error profile will focus on recall. On the Cleveland test set, the confusion matrix (Figure 6) shows that most patients with an angiogram, which uses contrast dye and X-rays to see the blood vessels in heart disease, are correctly flagged. A smaller subset of patients without heart disease is incorrectly predicted as positive. False negatives are rare at this threshold, which is intended to minimize missed high-risk cases in screening. Together with the ROC curve and threshold metrics, this shows a clear picture of how the model behaves at the selected cutoff.

Figure 6.



Interpretability is critical for clinicians to understand the data. Logistic regression coefficients were studied after preprocessing to identify the most influential predictors. The 10 features with the most significant absolute coefficients (Figure 7) align well with clinical explanations. Positive coefficients for features such as more typical angina patterns (cp_4.0), more abnormal thallium results (thal_7.0), the interaction of ST depression with exercise-induced angina (oldpeak_x_exang), male sex (sex_1.0), and the presence of at least one affected vessel (ca_1.0) indicate a higher estimated risk of heart disease. Negative coefficients for features such as no visible coronary narrowing (ca_0.0), specific chest pain categories (cp_3.0), and higher values of the fitness proxy thalach_per_age indicate lower risk. These directions and magnitudes are consistent with previous analyses of the Cleveland dataset and other heart-disease prediction studies [3], [7],[8]. This supports the idea that the model is capturing real risk patterns and achieving stronger results.

Figure 7.

To explore the portability, a reduced three-feature logistic model (age, sex, and systolic blood pressure) was trained on the Cleveland data and evaluated both internally and on the cleaned Kaggle cardiovascular dataset [3], [4]. As expected, performance decreased when using only three predictors and when transporting the model to a different population and schema. The ROC-AUC dropped to roughly 0.68 for the Cleveland dataset and 0.59 for Kaggle (Table 4). These results illustrate that a simple model can provide a sense of cardiovascular risk. Performance will degrade when applied to new settings without retraining or recalibration, or when only a limited subset of predictors is available.

Table 4: Reduced three-feature model: Internal vs External dataset performance.

Dataset	**ROC-AUC**	**Recall**	**Precision**	**F1**
Internal (Cleveland, reduced)	0.683	0.57	0.65	0.61
External (Kaggle, reduced)	0.593	0.26	0.61	0.36

At the selected threshold of 0.45, the calibrated logistic model met the pre-specified performance targets (test ROC-AUC ≈ 0.94, recall ≈ 0.89, precision ≈ 0.84), supporting its use as a high-recall screening tool for identifying patients at elevated risk of heart disease. These results support three main recommendations. First, a calibrated logistic regression model using a modest set of interpretable features can achieve strong discrimination and clinically sensible behavior on a population similar to that in the Cleveland dataset. Second, threshold selection should be treated as a policy decision and revisited with stakeholders. Metrics and confusion matrices are used to balance recall with the workload and costs associated with false positives. Third, external uses of the model should always include local validation and recalibration to reflect the characteristics of the new features and data sources [1],[4],[5].

6. Conclusion

The project developed and evaluated a supervised machine-learning model to estimate the probability of heart disease using routinely collected clinical variables. Using the Cleveland Heart Disease dataset as the primary source, the calibrated logistic regression model achieves strong results. This showed a ROC-

AUC of around 0.94 and high recall at a clinically reasonable operating threshold, with risk factors that align with established cardiovascular knowledge [2],[3],[7],[8].

The analysis showed that relatively simple, interpretable models can provide actionable risk and support early-detection strategies for heart disease. This model is deployed with a clear threshold and can be integrated easily into clinical workflows. The reduced-external experiment highlighted that portability is limited. Performance across new datasets illustrates the importance of local validation, recalibration, and monitoring.

Future work could extend this project by adding more data sources, such as medication histories, and further separating by race (Caucasian, Asian, African American, etc.). The other is to explore alternative modeling approaches and perform more comprehensive external validation across multiple healthcare datasets. Within the constraints of the available data, the project demonstrates that a calibrated, interpretable logistic model can help clinics prioritize patients for further cardiovascular evaluation. This shows that machine learning has the potential to reduce the burden of undetected heart disease.

References

[1] World Health Organization. (2025, July 31). *Cardiovascular diseases (CVDs) – key facts*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) World Health Organization+1

[2] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., … American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2023). Heart disease and stroke statistics—2023 update: A report from the American Heart Association. *Circulation, 147*(8), e93–e621. https://doi.org/10.1161/CIR.0000000000001123 AHA Journals+1

[3] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology, 64*(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9 PubMed+1

[4] Kaggle. (n.d.). *Cardiovascular disease dataset (cardio_train.csv)* [Data set]. Kaggle. Retrieved 2025 from https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset Kaggle+1

[5] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). Association for Computing Machinery. https://doi.org/10.1145/1102351.1102430 ACM Digital Library+1

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. http://www.jmlr.org/papers/v12/pedregosa11a.html Journal of Machine Learning Research+1

[7] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked, 19*, 100330. https://doi.org/10.1016/j.imu.2020.100330 ScienceDirect+1

[8] Osei-Nkwantabisa, Y., & Ntumy, R. (2024). Comparative analysis of machine learning algorithms for heart disease prediction using the UCI dataset. *Journal of Healthcare Informatics Research*. Advance online publication. https://doi.org/10.3233/HIS-240017 SAGE Journals