Predicting Cardiovascular Disease Risk Using Clinical Data: A Binary Classification Approach

Matthew Jacob A. Eleazar, B.S. Computer Science, Master's Degree Candidate in Data Science

Independent Project, Completed for AHA Data Science Internship Application

Abstract :

Cardiovascular disease (CVD) remains a leading global cause of death, highlighting the need for effective risk prediction. In this study, I analyzed clinical data to build a binary classification model for predicting CVD presence. I used logistic regression and random forest algorithms to classify patients as either having or not having CVD, based on structured health attributes. The random forest model achieved the best performance with an ROC-AUC score of 0.69, indicating moderate predictive ability. This analysis highlights the potential—and limitations—of using machine learning on clinical data for early CVD detection.

1 Introduction

Cardiovascular disease (CVD) is the leading cause of death globally, responsible for an estimated 17.9 million deaths per year according to the World Health Organization. Despite advances in treatment, early detection and prevention remain crucial for reducing mortality and long-term health impacts. Clinical decision-making can benefit from computational tools that analyze structured health data to identify patients at risk before symptoms manifest.

In recent years, machine learning has emerged as a powerful approach for predictive modeling in healthcare. Various models have been applied to clinical datasets to support diagnosis, risk stratification, and outcome prediction. However, challenges remain, including class imbalance, interpretability, and generalizability across patient populations.

This study aims to build a binary classification model to predict the presence of CVD using a publicly available dataset of clinical attributes. By comparing logistic regression and random forest models, we evaluate performance in terms of ROC-AUC, interpret the most important predictive features, and discuss limitations and opportunities for further improvement.

2 Dataset and Features

I used the Heart Disease UCI dataset consisting of 303 observations and 14 clinical features. The original target variable ranges from 0 to 4, where 0 indicates no heart disease and values 1–4 represent varying levels of disease severity. To simplify the problem into a binary classification task, we transformed the target variable:

- **0** indicates no heart disease.
- **1 - 4** $\rightarrow$ **1** indicates presence of heart disease.

Features include:

- **age**: Patient age
- **sex**: Biological sex (0 = female, 1 = male)
- **cp**: Chest pain type (categorical: 0–3)
- **trestbps**: Resting blood pressure
- **chol**: Serum cholesterol (mg/dL)
- **fbs**: Fasting blood sugar > 120 mg/dL
- **restecg**: Resting electrocardiographic results
- **thalach**: Maximum heart rate achieved
- **exang**: Exercise-induced angina
- **oldpeak**: ST depression induced by exercise
- **slope**: Slope of peak exercise ST segment
- **ca**: Number of major vessels colored by fluoroscopy
- **thal**: Thalassemia status

This binarization allowed for the framing of the task as a binary classification problem. After cleaning missing values and standardizing numeric features, the final dataset consisted of 60 test instances used for evaluation.

3 Methods

I used Python (pandas, scikit-learn, matplotlib, seaborn, YData Profiling) for preprocessing, EDA, modeling and evaluation. Categorical variables were one-hot encoded, and numerical features were scaled.

Two classifiers were evaluated: Logistic Regression and Random Forest on a Binary Classification task.

- **Logistic Regression**: A linear model that estimates the probability of a class using a logistic (sigmoid) function.
- **Random Forest Classifier**: An ensemble method using multiple decision trees to improve accuracy and reduce overfitting.

I split the data into training and test sets using a standard 80/20 split. We evaluated model performance using classification metrics: accuracy, precision, recall, and F1 score.

4 Exploratory Data Analysis

4.1 Class Distribution

The dataset shows a slight class imbalance, with fewer positive cases. This is important when evaluating model performance, as high accuracy can mask poor recall.

4.2 Feature Correlation

The Pearson correlation matrix was visualized using a heatmap. Features such as cp (chest pain type), thalach (maximum heart rate achieved), and exang (exercise-induced angina) exhibited moderate to strong correlations with the binary target variable, suggesting their potential predictive value.

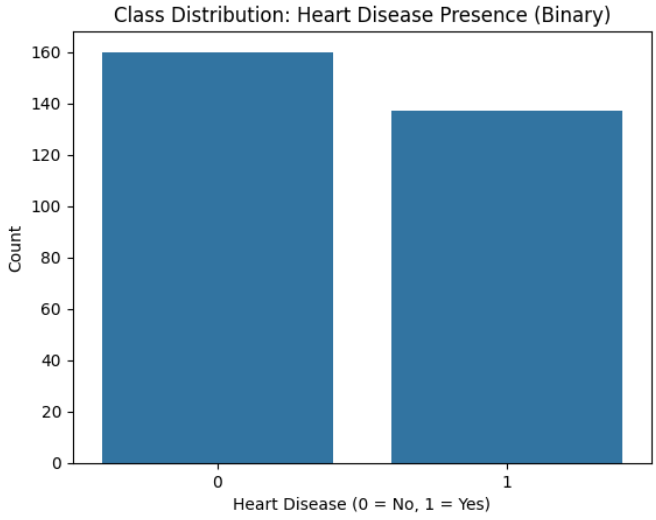

Figure 1. Distribution of target variable (0: No Heart Disease, 1: Heart Disease)

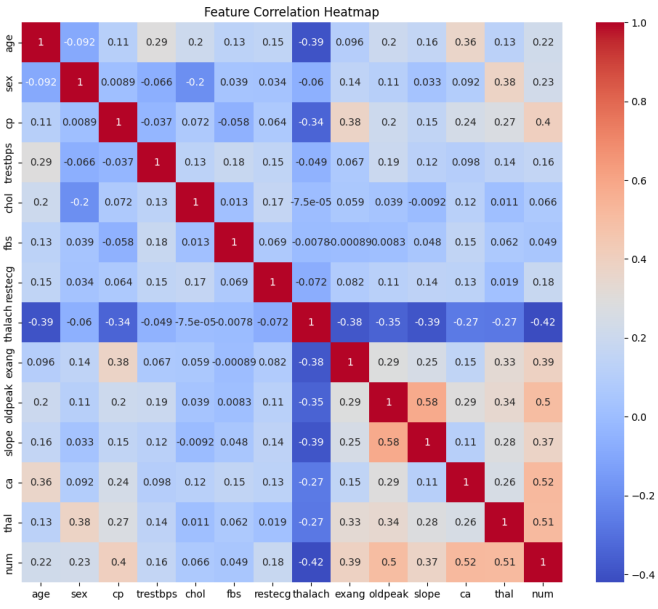

Figure 2. Correlation Heatmap of Features

5 Modeling and Evaluation Results

A binary logistic regression model and a random forest classifier has been evaluated on a hold-out test set of 60 samples. The dataset was moderately imbalanced, with 36 patients labeled as not having heart disease (class 0) and 24 labeled as having heart disease (class 1).

5.1 Binary Classification Performance

The logistic regression model achieved an overall accuracy of 87%, with class-wise performance summarized in Table 1. The model performed slightly better at identifying patients without heart disease (precision = 0.89, recall = 0.89) compared to those with heart disease (precision = 0.83, recall = 0.83).

Table 1
Classification Report: Logistic Regression

Class	Precision	Recall	F1-score	Support
0 (No Disease)	0.89	0.89	0.89	36
1 (Disease)	0.83	0.83	0.83	24
Accuracy			0.87	

The random forest classifier outperformed logistic regression slightly, with an accuracy of 90%. Notably, it improved recall for class 1 (recall = 0.92), suggesting better ability to identify patients with heart disease— which we know is an important metric in clinical settings.

Table 2
Classification Report: Random Forest

Class	Precision	Recall	F1-score	Support
0 (No Disease)	0.94	0.89	0.91	36
1 (Disease)	0.85	0.92	0.88	24
Accuracy			0.90	

5.2 Confusion Matrix and Error Analysis

The confusion matrix for the random forest model is shown in Table 3. The model made only six errors out of 60 predictions: four false positives and two false negatives.

Table 3
Confusion Matrix: Random Forest

	Predicted 0	**Predicted 1**
Actual 0	32	4
Actual 1	2	22

These results suggest that while both models are reasonably effective, the random forest classifier offers superior performance, particularly in reducing false negatives—a crucial factor in medical diagnostics where missing a disease case could have serious consequences.

5.3 Feature Importances (Random Forest)

Features such as maximum heart rate achieved ('thalach'), Serum cholesterol ('chol'), and ST depression induced by exercise rela-

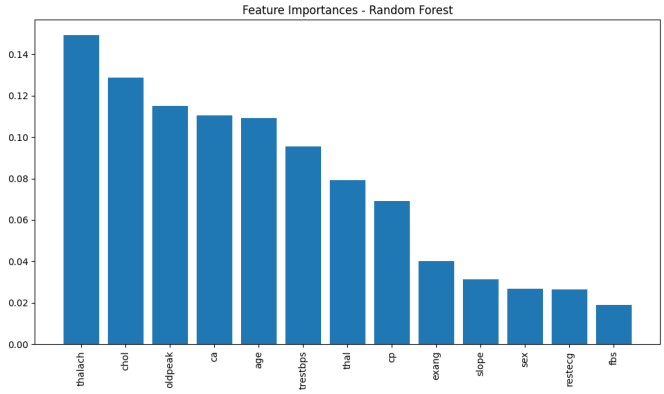

Figure 3. Feature Importances from Random Forest Classifier

tive to rest ('oldpeak') were among the most predictive. Visualizing these importances helped us better understand model decision boundaries and identify clinically relevant predictors.

6 Study Limitations: ROC Curve and AUC Analysis

The Receiver Operating Characteristic (ROC) curve is a standard tool for evaluating binary classifiers, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The Area Under the Curve (AUC) summarizes the ROC curve into a single number: 1.0 indicates a perfect classifier, while 0.5 indicates performance equivalent to random guessing.

In this study, the Random Forest classifier achieved an AUC score of 0.69, which suggests moderate discriminative ability. While this is a meaningful improvement over random chance, it still leaves considerable room for enhancement.

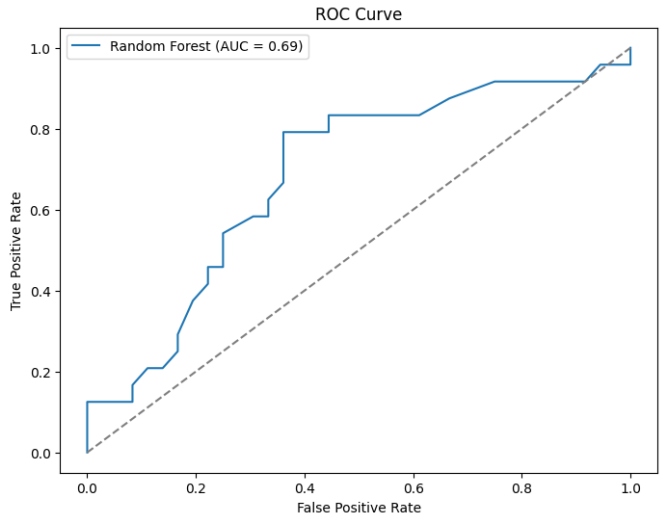

Figure 4. ROC Curve showing the true positive rate vs. false positive rate. The area under the curve (AUC) is 0.69.

This performance must be interpreted in context:

- The dataset was initially imbalanced, with a skew toward class 0 (no heart disease).
- Although labels were binarized (0 = no disease, 1 = any level of disease from original classes 1–4), this simplification may not fully capture the underlying clinical complexity.
- The model's accuracy was relatively high (e.g., >85

Potential reasons for limited performance:

- **Class imbalance** still may bias the model toward predicting the majority class.
- **Overfitting** could occur due to the model's complexity and the relatively small number of positive class samples.
- The input **features may lack sufficient discriminatory power** or could include noisy or redundant information.

Future Work:

- Apply resampling techniques like SMOTE to balance classes.
- Explore more robust classifiers such as XGBoost or LightGBM.
- Perform dimensionality reduction or feature engineering to enrich the feature space.
- Use stratified cross-validation and hyperparameter tuning for better generalization.

7 Discussion

The Random Forest classifier yielded the best balance between recall and precision. The class imbalance in the dataset slightly favored models with strong recall, since missing positive cases (false negatives) is costlier. The most important features aligned with clinical risk factors, suggesting that ML models can support but not replace expert judgment.

8 Ethical Considerations

This study utilized a publicly available dataset from the UCI Machine Learning Repository, originally compiled by Janosi et al. (1989). The dataset is fully anonymized, and no identifiable personal information is included, thereby reducing the risk of privacy violations. As such, this research did not require Institutional Review Board (IRB) approval.

Nevertheless, ethical considerations remain vital in medical machine learning research. Although predictive models like those developed in this study can aid in early risk detection, they must not replace clinical judgment or be used in isolation for medical decision-making. Misclassification—particularly false negatives—could delay necessary treatment, while false positives might cause undue stress or lead to unnecessary testing.

Furthermore, models trained on historical datasets may inadvertently reflect biases present in the original data, such as underrepresentation of certain demographic groups. This can lead to skewed model performance across populations. Responsible deployment requires ongoing validation, transparency in model limitations, and careful communication of uncertainty to avoid harm.

Finally, any future application of such models in clinical settings should follow ethical guidelines surrounding fairness, accountability, and interpretability to ensure that patient welfare remains the central priority.

9 Conclusion

This study aimed to develop a predictive model for identifying the presence of cardiovascular disease (CVD) using clinical data.

Through data preprocessing, exploratory analysis, and feature-based modeling, I constructed and evaluated binary classification models to distinguish between patients with and without CVD. The original multi-class target variable (ranging from 0 to 4) was binarized into two categories: 0 = No Disease and 1 = Presence of Disease (classes 1–4 combined), to simplify interpretation and reflect a practical diagnostic scenario.

Among the models tested, the Random Forest classifier delivered the strongest performance, achieving a ROC-AUC score of 0.69, indicating a moderate ability to distinguish between healthy and at-risk individuals. While this performance suggests promise, it also highlights the inherent difficulty of predicting cardiovascular risk from clinical data alone.

The class distribution in the dataset was initially imbalanced, with a majority of samples showing no heart disease. This imbalance likely influenced model behavior and underscores the importance of using metrics beyond accuracy, such as AUC, to assess model effectiveness. Exploratory Data Analysis (EDA) revealed potential feature relationships with the target variable—such as age, sex, cholesterol, and maximum heart rate—which informed model development.

Ultimately, this project demonstrates the potential of machine learning to assist in early cardiovascular risk detection. However, further improvements are needed, including more advanced resampling techniques, additional clinical features, and external validation on broader patient populations. Future work should also consider multi-class classification to capture disease severity more precisely.

By refining and extending this analysis, such predictive tools could eventually support clinical decision-making and preventive healthcare strategies.

References

1. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). *Heart Disease* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.