Interpretable Machine Learning for Cardiovascular Risk Stratification

Saravana Priyaa C R

1 Introduction

Cardiovascular disease remains a leading cause of mortality worldwide, making early risk identification a critical component of preventive care. In this project, I use structured, de-identified clinical data provided for the challenge to develop interpretable machine learning models for heart disease risk prediction.

Rather than optimizing solely for accuracy, the emphasis is placed on clinically meaningful performance, model transparency, and practical risk stratification.The goal is to support early screening and triage decisions by producing reliable risk estimates that clinicians can reasonably interpret and trust.

2 Exploratory Data Analysis

2.1 Data Overview and Quality

The analyzed dataset contains 918 patient records with 16 features, including a binary target variable indicating the presence of heart disease. All features are numerically encoded, enabling direct statistical analysis without additional preprocessing. A completeness check confirmed that no missing values were present, allowing the analysis to focus on feature behavior rather than data repair.

2.2 Target Distribution

The target variable is relatively balanced, with 55.3% positive and 44.7% negative cases. This reduces the risk of strong class bias and supports the use of standard classification metrics.

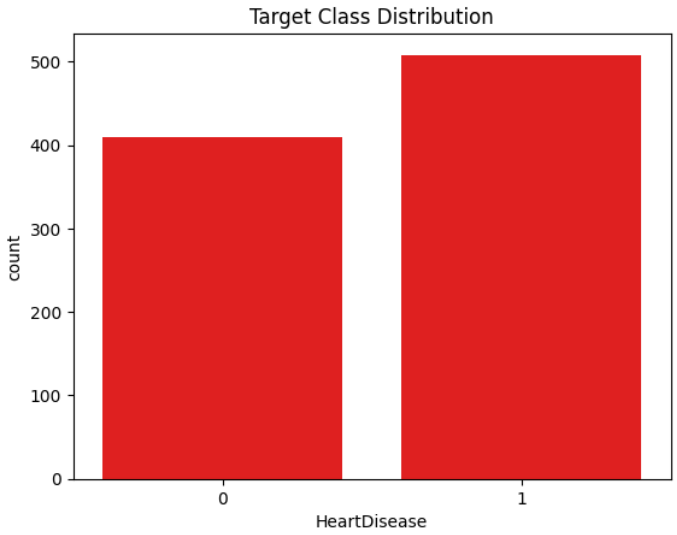

Figure 1: Target class distribution

2.3 Feature Correlation Analysis

Pairwise correlations between features and the target outcome were examined to identify clinically relevant predictors. The strongest associations were observed for ECG- and exercise-related features:

- **ST_Slope_Flat** showed a strong positive correlation with heart disease.
- **ST_Slope_Up** exhibited a strong negative correlation, suggesting a protective relationship.
- **Exercise-induced angina** and **ST depression (Oldpeak)** were positively associated with increased risk.

Other variables such as maximum heart rate, age, and fasting blood sugar showed moderate correlations, indicating supporting rather than dominant roles.

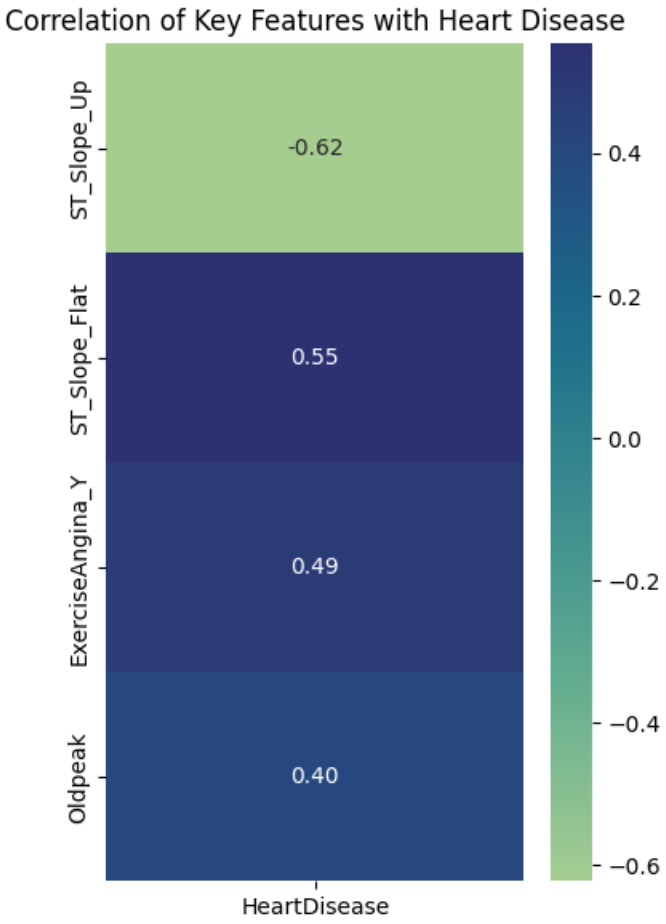

Figure 2: Top feature correlations with heart disease

Overall, these patterns align well with established clinical understanding, suggesting that the dataset captures meaningful and interpretable cardiovascular signals.

3 Modeling Approach

3.1 Baseline Model: Logistic Regression

A logistic regression model was trained as a transparent baseline to assess whether linear relationships were sufficient for risk prediction. On a held-out test set, the model achieved a ROC-AUC of 0.93 with high recall (0.93) for the heart disease class. The confusion matrix showed relatively few false negatives, which is desirable for early screening scenarios.

Coefficient analysis revealed clinically intuitive relationships: flat ST slope, exercise-induced angina, elevated fasting blood sugar, and higher Oldpeak values increased predicted risk, while upward ST slope and higher maximum heart rate were associated with lower risk.

3.2 Random Forest Model

To capture non-linear interactions while maintaining interpretability, a Random Forest classifier was trained using the same data split. The model achieved a ROC-AUC of 0.93, marginally improving upon the baseline while maintaining strong recall. Feature importance analysis highlighted ST-segment slope, Oldpeak, maximum heart rate, and exercise-induced angina as the most influential predictors.

4 Model Interpretability

Because high predictive performance alone is insufficient in clinical settings, SHAP (SHapley Additive exPlanations) was applied to verify that the Random Forest was learning clinically reasonable patterns rather than spurious correlations.

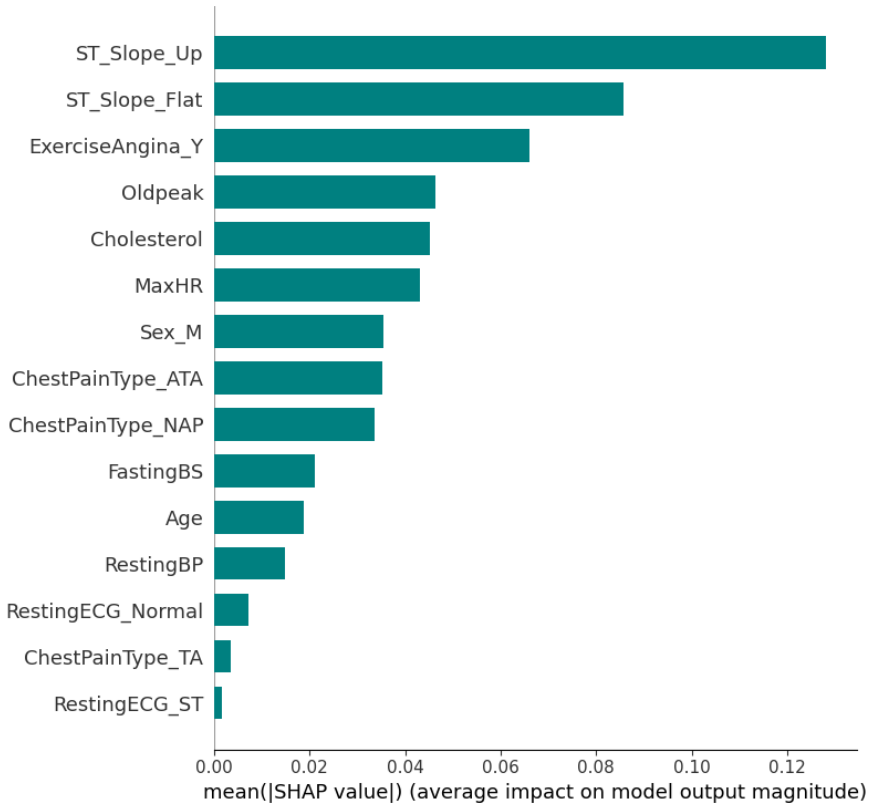

Figure 3: Global SHAP feature importance

The direction and magnitude of SHAP values aligned closely with both correlation analysis and model feature importance rankings, increasing trust in the model and supporting its use as a decision-support tool rather than a black-box classifier.

5 Risk Stratification and Calibration

5.1 Risk Stratification

Predicted probabilities from the Random Forest model were used to stratify individuals into low, medium, and high cardiovascular risk groups. Clear separation was observed: over 90% of individuals in the high-risk group had heart disease, while the low-risk group was predominantly disease-free.

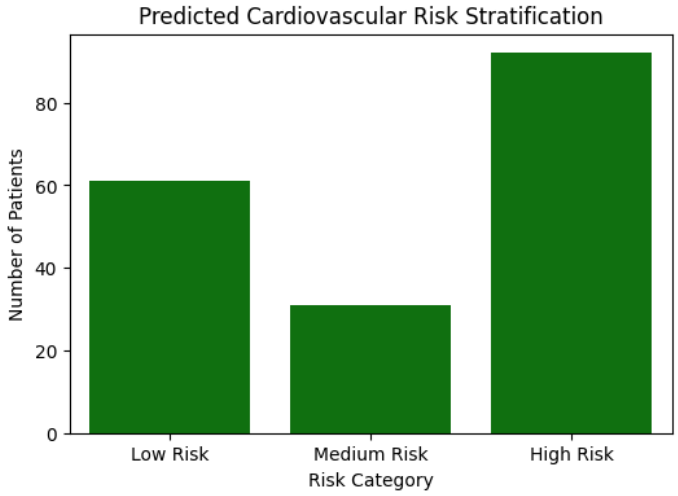

Figure 4: Predicted cardiovascular risk stratification

5.2 Probability Calibration

A calibration analysis demonstrated close agreement between predicted probabilities and observed outcomes, particularly in the mid-to-high risk range most relevant for screening. The Brier score further indicated reasonable probability calibration.

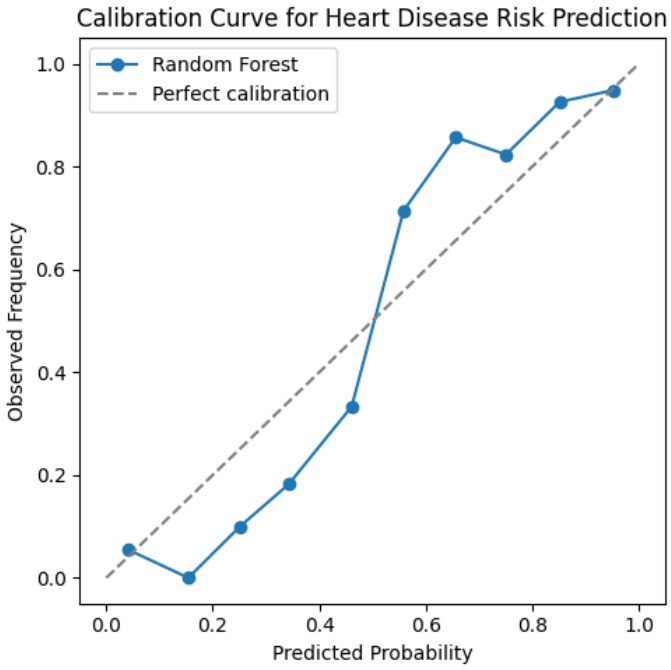

Figure 5: Calibration curve for heart disease risk prediction

6 Limitations and Future Work

This study focuses on structured clinical features to prioritize interpretability within the scope of the challenge. Raw physiological signals such as ECG time-series were not included but represent a promising direction for future work. Additionally, evaluation was performed using a single train–test split; future work could include cross-dataset validation across other provided datasets to assess robustness and generalizability.