Comparative Analysis of Classification Methods for Heart Disease Prediction

Abstract

Heart disease is one of the leading causes of morbidity and mortality worldwide and continues to pose a significant burden on healthcare systems. Early identification of individuals at high risk of heart disease is therefore of considerable importance, as timely intervention and preventive strategies can improve patient outcomes and reduce healthcare costs. In this project, the problem of heart disease prediction was investigated using routinely collected clinical and demographic data, with the aim of comparing the performance of traditional statistical modeling approaches and modern machine learning methods.

A publicly available heart disease dataset, containing clinical records from 918 patients, was analyzed within the R statistical computing environment. Exploratory data analysis was conducted to examine variable distributions, assess class imbalance, and explore relationships between predictors and heart disease status. Two classification models, logistic regression and random forest, were subsequently implemented to predict the presence of heart disease. Logistic regression was used as a baseline model due to its interpretability and widespread use in clinical research, while random forest was employed to capture potential nonlinear relationships and interactions among predictors.

Model performance was evaluated using classification accuracy and the area under the receiver operating characteristic curve (AUC). In addition, a simulation-based framework was applied to examine how predictive performance changes with increasing sample size and to assess model stability. The results indicate that both models are capable of distinguishing between patients with and without heart disease; however, random forest consistently achieves higher predictive accuracy and AUC across all evaluated sample sizes. Logistic regression, while less accurate, demonstrates more stable performance and greater interpretability. These findings highlight an important trade-off between predictive performance and model transparency in the context of heart disease prediction.

1. Introduction

Heart failure is a major global public health concern and represents one of the leading causes of morbidity and mortality worldwide. According to the World Heart Federation, more than 64 million people are affected by heart failure globally, placing a substantial burden on patients, healthcare systems, and economies. The condition is often chronic and progressive, leading to frequent hospitalizations, reduced quality of life, and increased healthcare costs. Early identification of individuals at high risk of heart failure is therefore essential, as timely intervention and preventive strategies can significantly improve clinical outcomes and reduce disease progression.

In recent years, the widespread availability of clinical and demographic data, together with advances in statistical computing, has enabled the use of data-driven approaches for disease prediction. Traditional statistical models and modern machine learning techniques are increasingly applied in healthcare to support clinical decision-making. These methods can identify complex patterns and relationships among risk factors that may not be apparent through conventional analysis alone. As a result, predictive modeling has become an important tool for assessing heart failure risk and prioritizing patients for further clinical evaluation.

Among the various classification techniques used in medical prediction tasks, logistic regression and random forest are two of the most applied methods. Logistic regression has long been favored in clinical research due to its simplicity, interpretability, and solid theoretical foundation. It allows clinicians and researchers to quantify the association between risk factors such as age, sex, blood pressure, cholesterol levels, and exercise-induced symptoms and the probability of heart failure. However, logistic regression relies on linear assumptions that may limit its ability to capture complex nonlinear relationships inherent in physiological data.

In contrast, random forest is a powerful ensemble learning method that combines multiple decision trees to achieve high predictive accuracy and robustness. By incorporating randomness in both data sampling and feature selection, random forests can model nonlinear effects and interactions between predictors while reducing the risk of overfitting. Previous studies have shown that random forest models often outperform traditional statistical methods in classification tasks involving heterogeneous clinical data. Nevertheless, their performance can vary depending on the dataset characteristics, feature selection, and sample size, highlighting the need for systematic comparison with more interpretable models.

As machine learning methods become increasingly integrated into healthcare applications, it is important to evaluate and compare commonly used algorithms to determine their suitability for heart failure prediction. Publicly available datasets provide an opportunity to conduct transparent and reproducible comparisons across different modeling approaches. In this study, a heart failure prediction dataset obtained from Kaggle is used, consisting of clinical and demographic records from 918 patients. The dataset includes key variables such as age, sex, chest pain type, resting

blood pressure, cholesterol levels, electrocardiographic results, and exercise-related indicators, all of which are relevant to cardiovascular risk assessment.

The primary aim of this study is to perform a comparative analysis of logistic regression and random forest models for heart failure prediction. Model performance is evaluated using multiple metrics, including accuracy and the area under the receiver operating characteristic curve (ROC-AUC), to provide a comprehensive assessment of classification ability. In addition, feature selection techniques are employed to identify the most influential clinical and demographic predictors, thereby enhancing model efficiency and interpretability. A simulation-based approach is further used to investigate how predictive performance changes with increasing sample size, offering insight into model stability and generalization.

By addressing these objectives, this research aims to contribute to a better understanding of the strengths and limitations of statistical and machine learning methods in heart failure prediction, and to provide guidance for their application in healthcare settings.

2. Data

2.1 Data Source and Description

The dataset used in this study is the *Heart Failure Prediction* dataset obtained from Kaggle. It consists of 918 observations, where each observation corresponds to a single patient. The response variable, 'HeartDisease', is binary and indicates whether the patient has been diagnosed with heart disease (1 = Yes, 0 = No). The dataset includes a mix of numerical and categorical predictors that are clinically relevant to cardiovascular health.

The main variables in the dataset are summarized as follows:

- Age: Age of the patient in years (numeric).
- Sex: Biological sex of the patient (Male/Female).
- ChestPainType: Type of chest pain experienced (TA, ATA, NAP, ASY).
- RestingBP: Resting blood pressure in mm Hg (numeric).
- Cholesterol: Serum cholesterol level in mg/dl (numeric).
- FastingBS: Fasting blood sugar (>120 mg/dl: 1 = true, 0 = false).
- RestingECG: Resting electrocardiographic results (Normal, ST, LVH).
- MaxHR: Maximum heart rate achieved during exercise (numeric).
- ExerciseAngina: Exercise-induced angina (Yes/No).
- Oldpeak: ST depression induced by exercise relative to rest (numeric).
- ST_Slope: Slope of the peak exercise ST segment (Up, Flat, Down).
- HeartDisease: Target variable indicating heart disease status.

A preliminary data quality check showed that the dataset contains no missing values across all variables, making it suitable for direct modeling after appropriate preprocessing.

2.2 Data Preprocessing

All analyses were conducted in R. Categorical variables were converted into factor variables with clinically interpretable labels. The response variable 'HeartDisease' was encoded as a two-level factor (No, Yes) to ensure compatibility with classification models in the caret framework.

Given the absence of missing values, no imputation was required. However, careful attention was paid to factor level ordering and labeling to avoid modeling inconsistencies.

2.3 Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to understand the structure of the dataset, assess class imbalance, examine distributions of variables, and explore relationships between predictors and the target variable. The EDA results are summarized using figures and tables, accompanied by interpretations.

2.3.1 Class Distribution of the Target Variable

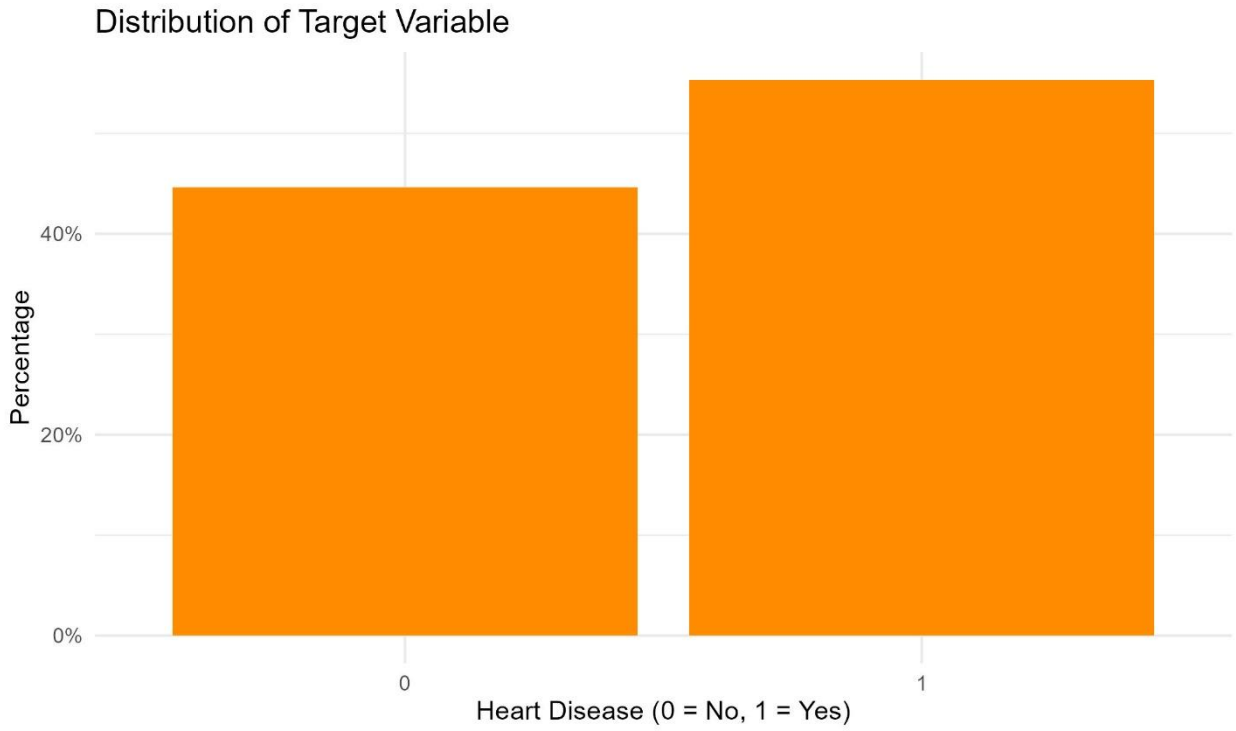


Figure 2.1 Proportion of patients with and without heart disease.

The bar chart shows that the dataset is moderately imbalanced, with a larger proportion of observations classified as having heart disease.

2.3.2 Descriptive Statistics of Numerical Variables

Summary statistics for numerical variables are presented in Table 2.1. These include measures of central tendency and dispersion for each continuous predictor.

Table 2.1: Summary Statistics of Numerical Variables

Variable	Min	Median	Mean	Max
Age	28	54	53	77
RestingBP	0.0	130.0	132.4	200.0

Cholesterol	0.0	223.0	198.8	603.0
MaxHR	60.0	138.0	136.8	202.0
Oldpeak	-2.6	0.6	0.88	6.2
Fasting Blood Sugar	0.0	0.0	0.23	1.0

According to the Table 2.1, Age is concentrated in middle-aged and older individuals, reflecting the higher prevalence of heart disease in these groups. Cholesterol and Oldpeak show high variability and right-skewness, indicating the presence of extreme values that may influence model fitting.

2.3.3 Distribution of Numerical Variables

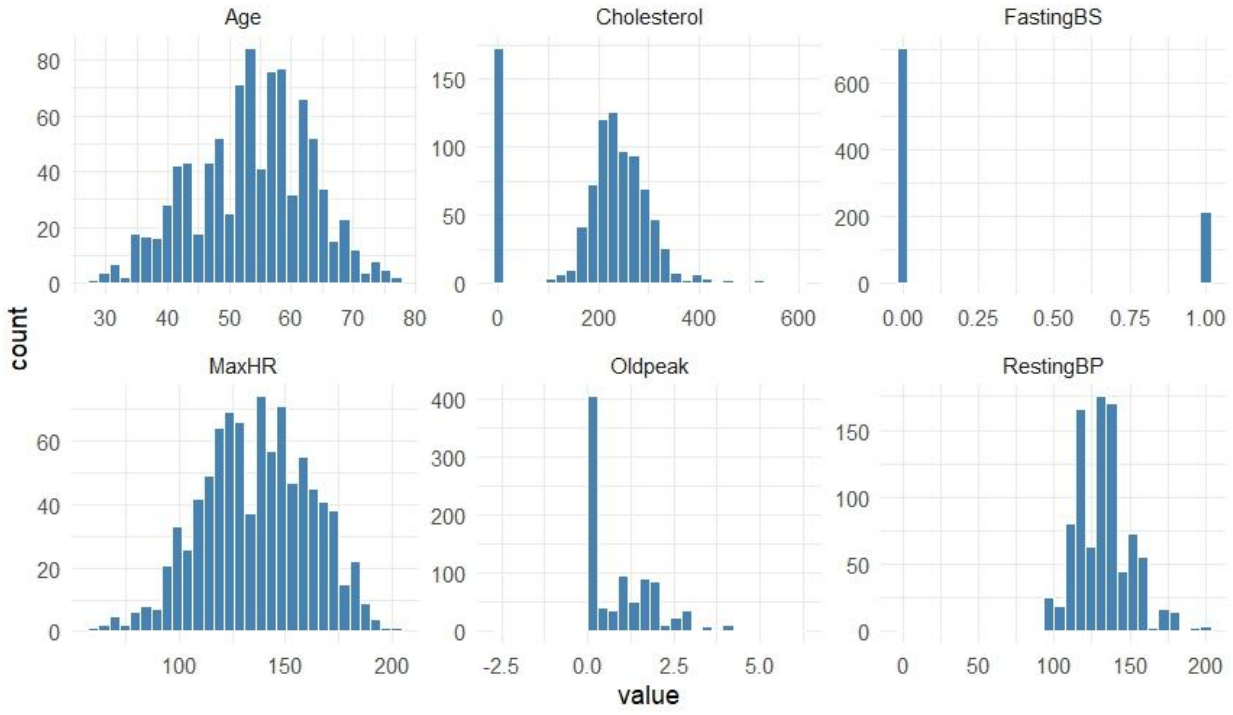


Figure 2.2 Distribution of Numerical Variables

Figure 2.2 presents histograms of all numerical predictors. The plots reveal that several variables deviate from normality. Age and Max Heart Rate exhibit approximately symmetric distributions, while Cholesterol and Oldpeak are strongly right-skewed. These distributional characteristics motivate the use of flexible, non-parametric models such as random forests alongside parametric models like logistic regression.

2.3.4 Numerical Variables by Heart Disease Status

Boxplots comparing numerical variables across heart disease status are shown in Figure 2.3.

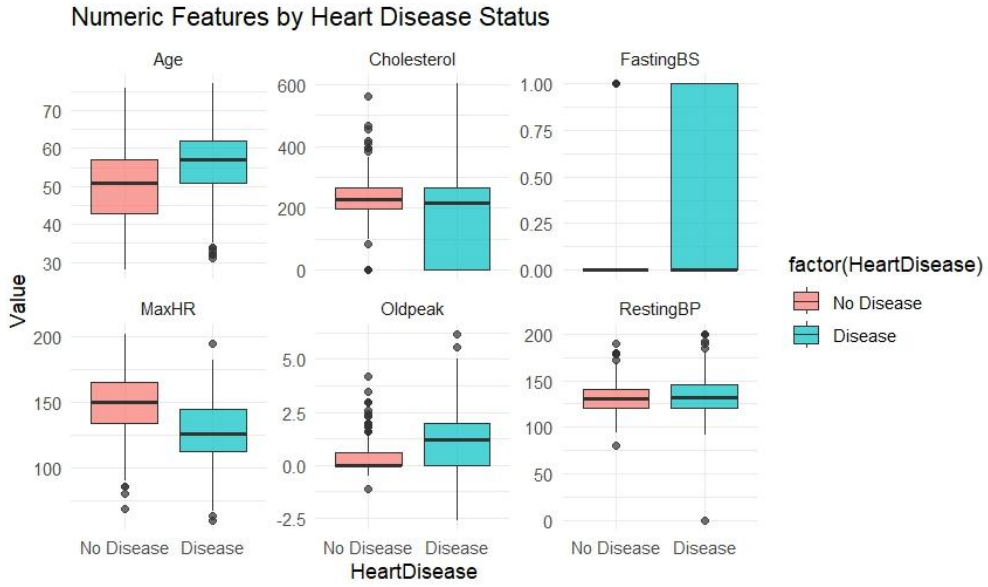


Figure 2.3 Numerical variables vs Heart Disease Status

Patients diagnosed with heart disease tend to have lower maximum heart rates and higher Oldpeak values, suggesting reduced cardiovascular fitness and more severe exercise-induced ST depression. These clear separations indicate strong discriminatory power for these variables.

2.3.5 Distribution of Categorical Variables

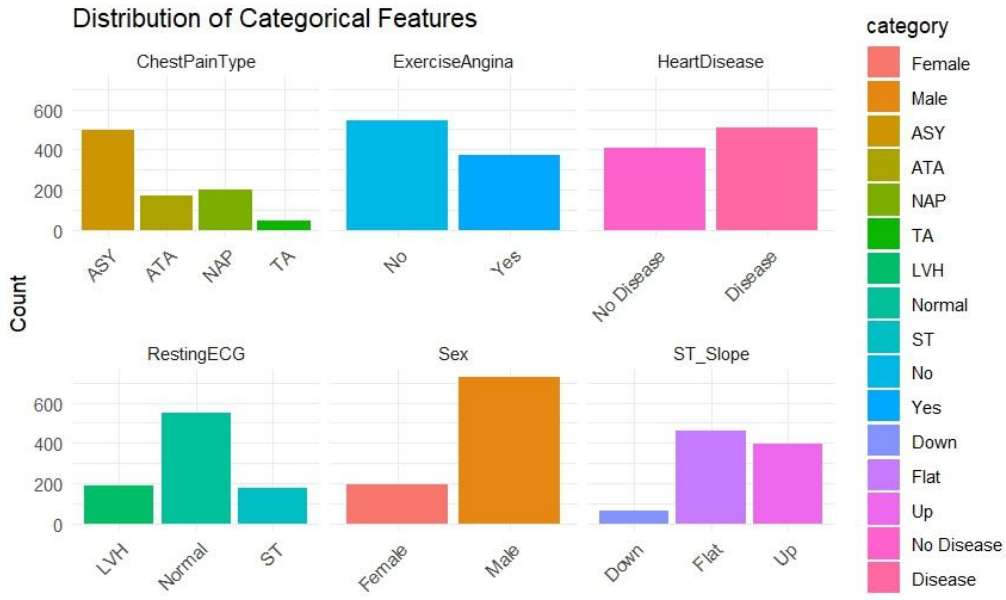
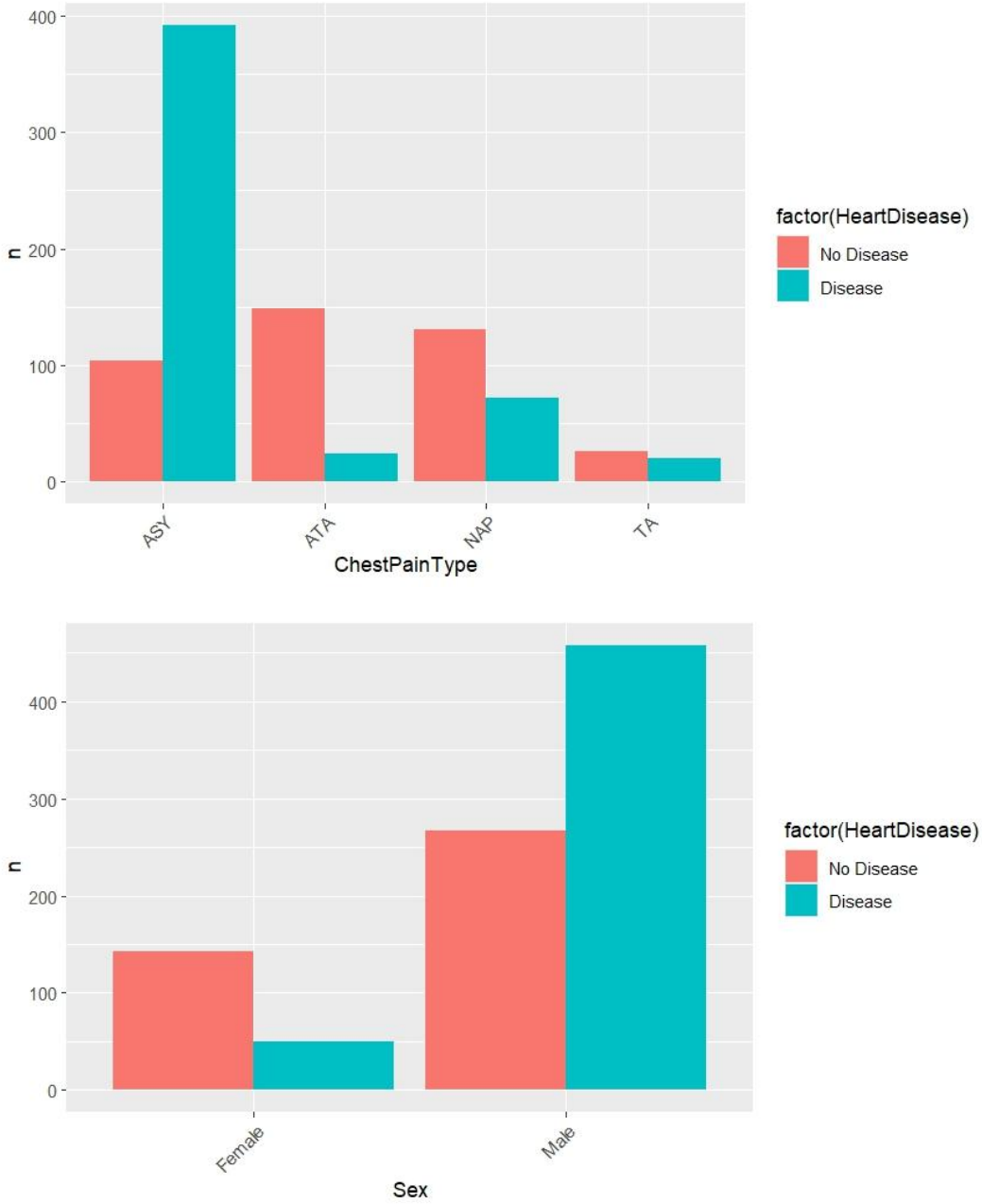


Figure 2.4 Distribution of Categorical Variables

The figure shows that the dataset is dominated by male patients and that asymptomatic chest pain is the most common chest pain type, indicating that many patients with heart disease do not show typical symptoms. Exercise-induced angina and abnormal ST-segment slopes (flat or down) are common, both of which are clinically associated with higher cardiovascular risk. Overall, these distributions reveal strong and clinically meaningful patterns that support the use of predictive modeling for heart disease.

2.3.6 Categorical Variables vs Heart Disease Status



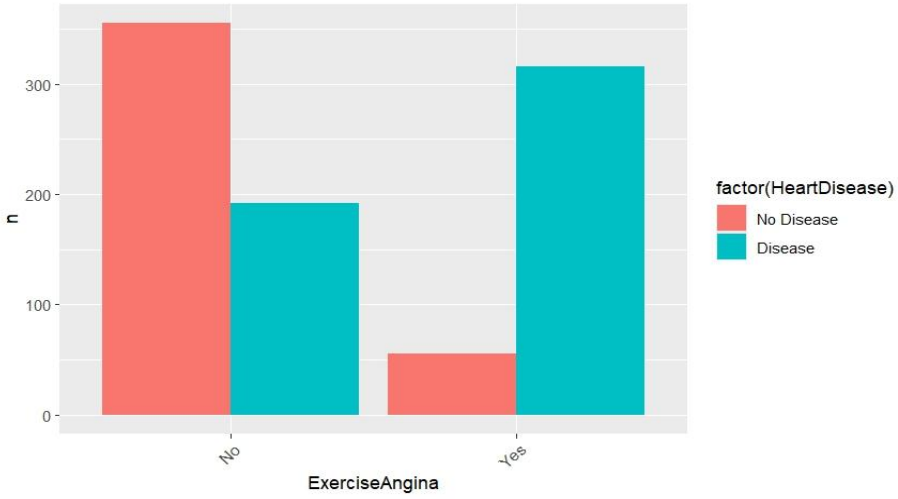


Figure 2.5 Categorical variables vs Heart Disease Status

Bivariate bar plots of selected categorical variables against heart disease status are shown in Figure 2.5. The proportion of heart disease cases is substantially higher among males and among patients experiencing exercise-induced angina. Chest pain type shows strong separation between disease classes, reinforcing its importance as a predictor.

2.3.7 Correlation Analysis

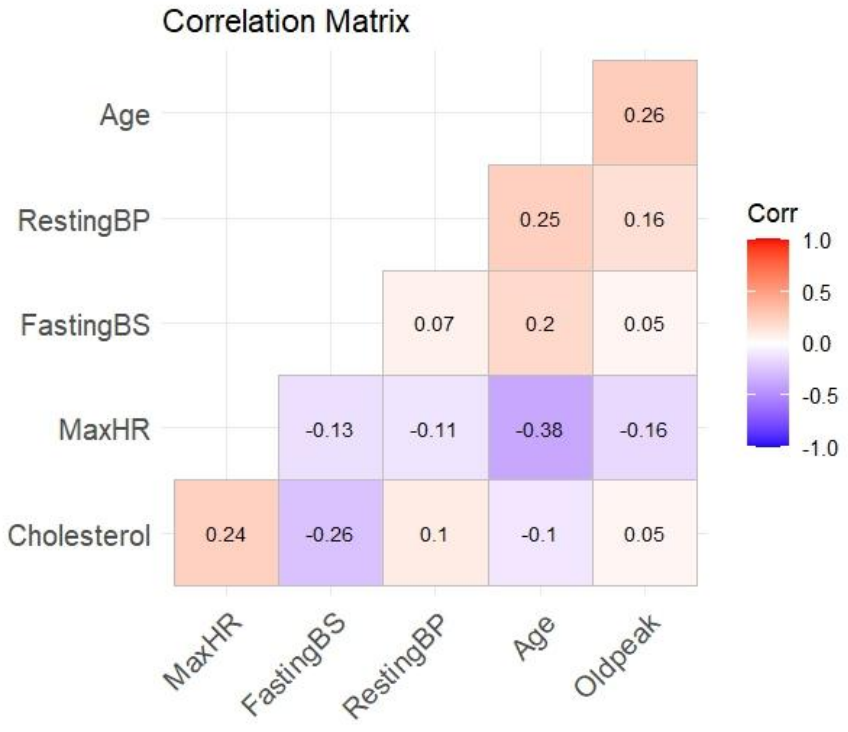


Figure 2.6 Correlation Matrix

The correlation matrix for numerical variables is shown in Figure 2.6.

The correlation matrix shows that Age and Oldpeak have a moderate positive correlation, indicating that older patients tend to have higher ST-segment depression. Max Heart Rate is negatively correlated with Age and Oldpeak, meaning that older or more clinically severe patients generally achieve lower maximum heart rates. Cholesterol and fasting blood sugar show only weak correlations with most variables, suggesting they provide independent information. Overall, no very strong correlations are present, indicating a low risk of multicollinearity in the model.

Overall, the EDA reveals clear and clinically interpretable differences between patients with and without heart disease. Both numerical and categorical variables demonstrate strong associations with the target variable, providing a solid foundation for the feature selection and predictive modeling approaches described in the Methods chapter.

3. Methods

3.1 Simulation Design

To investigate the effect of sample size on model performance, a simulation-based approach was adopted. Larger datasets were generated by sampling observations from the original dataset with replacement. Sample sizes of 1,000, 3,000, 5,000, and 10,000 were considered. For each simulated dataset, the same modeling pipeline was applied.

Each dataset was split into training (70%) and testing (30%) subsets using stratified sampling to preserve the class distribution of the target variable.

3.2 Feature Selection Using Recursive Feature Elimination

Recursive Feature Elimination (RFE) was employed on the training data to identify the most informative subset of predictors. RFE iteratively fits a model, ranks predictors by importance, and removes the least important features until an optimal subset is identified. In this study, RFE was implemented using random forest based variable importance measures with 5-fold cross-validation.

The optimal feature set identified by RFE was subsequently used for training both classification models, ensuring a fair comparison and reducing the risk of overfitting.

3.3 Classification Models

Two classification models were considered in this study: logistic regression and random forest. These models represent complementary approaches, combining interpretability from classical statistical modeling with the flexibility and predictive power of ensemble machine learning techniques.

3.3.1 Logistic Regression

Logistic regression is a widely used statistical model for binary classification problems, particularly in medical and epidemiological research.

Logistic regression provides a transparent baseline model, allowing direct interpretation of clinical risk factors. However, it assumes linearity in the log-odds, which may limit its ability to capture nonlinear physiological relationships.

3.3.2 Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and robustness. Each tree is trained on a bootstrap sample of the data, and at each split, a random subset of predictors is considered.

Random forests capture nonlinear relationships and interactions among predictors, are robust to multicollinearity and outliers, and provide variable importance measures used in feature selection.

Both models were trained using the caret package with 5-fold cross-validation. Class probabilities were estimated, and the ROC-based performance metric was used during training.

3.4 Model Evaluation Metrics

Model performance was evaluated on the held-out test set using:

- **Accuracy**: Proportion of correctly classified observations.
- **Area Under the ROC Curve (AUC)**: Measures the model's ability to discriminate between patients with and without heart disease across different classification thresholds.

Performance metrics were summarized across different simulated sample sizes, and marginal gains in accuracy were computed to assess the incremental benefit of increasing sample size.

This methodological framework enables a robust comparison of models while accounting for feature selection and sample size effects, providing insights into both predictive performance and practical modeling considerations in heart disease prediction.

4. Implementation

This chapter describes the implementation of the simulation study conducted to evaluate the performance of two classification models for heart disease prediction. The implementation was carried out entirely in the R statistical computing environment. The chapter outlines the R packages used, the simulation design, model specifications, and the performance metrics applied for model evaluation.

4.1 R Packages Used

The primary package used in this study is the Classification and Regression Training (caret) package. Caret provides a unified framework for data preprocessing, model training, hyperparameter tuning, resampling, and performance evaluation. Its standardized workflow enables fair and consistent comparison of different classification models under identical training and validation settings.

In this study, caret was used to train, tune, and evaluate all classification models. The most important caret functions applied include:

- **train()**: Used to fit classification models while integrating preprocessing, cross-validation, and hyperparameter tuning.
- **trainControl()**: Specifies the resampling strategy, including k-fold cross-validation and estimation of class probabilities.
- **createDataPartition()**: Splits the data into training and testing sets while preserving the class distribution of the outcome variable.
- **confusionMatrix()**: Computes classification accuracy and related performance measures.
- **rfe()**: Implements Recursive Feature Elimination (RFE), a wrapper-based feature selection method that iteratively removes less important predictors.
- **rfeControl()**: Defines control parameters for RFE, including cross-validation settings and the method used to rank predictor importance.

The integration of RFE within the caret framework allows feature selection to be embedded within the resampling process, reducing the risk of information leakage and overfitting. Using Random Forest based importance measures further ensures robustness to nonlinear relationships and interactions among predictors.

Several additional R packages were also used to support the analysis:

- **tidyverse**: Used for data cleaning, manipulation, and visualization.
- **randomForest**: Provides the Random Forest algorithm used within caret.

- **pROC**: Used to compute the Area Under the Receiver Operating Characteristic Curve (AUC).

4.2 Simulation Framework

The simulation study was designed to assess model performance under varying sample sizes. Subsamples of increasing size were repeatedly drawn from the full dataset. This approach allows investigation of how classification performance changes as more data become available, which is particularly relevant in medical research settings.

For each sample size, the following steps were performed:

1. A subset of data was selected from the full dataset.
2. The data were split into training and testing sets using stratified sampling.
3. Models were trained using k-fold cross-validation.
4. Performance metrics were computed on the test set.
5. Results were stored and summarized for comparison.

Two classification models were evaluated in this study:

- **Logistic Regression**: A baseline statistical model commonly used in clinical research due to its simplicity and interpretability.
- **Random Forest**: An ensemble learning method capable of capturing complex, nonlinear relationships among predictors.

4.3 Feature Selection

Feature selection is an important aspect of predictive modeling, particularly in clinical applications where predictors may be correlated or weakly informative. Identifying the most relevant features can improve model performance, interpretability, and clinical relevance.

In this study, feature selection was performed using embedded, model-based methods within the caret framework. Rather than applying feature elimination before model training, predictor importance was assessed during the resampling process. This approach helps prevent information leakage and ensures that feature selection is properly integrated with cross-validation.

Recursive Feature Elimination (RFE) combined with Random Forest was used to identify subsets of predictors that maximize predictive performance.

4.4 Simulation Functions

To ensure a systematic and reproducible simulation study, two custom R functions; **simulate_data()** and **run_simulation()** were developed. These functions automate data sampling, model training, and performance evaluation across different sample sizes.

- **simulate_data()**

The **simulate_data()** function generates datasets of varying sample sizes from the original heart disease dataset. Given the full dataset and a specified sample size, the function randomly selects a subset of observations, typically without replacement, while preserving the original data structure.

The outcome variable is retained as a factor to ensure compatibility with classification models in caret. Identical preprocessing steps are applied to each sampled dataset to ensure consistency across simulation runs. By repeatedly generating datasets of different sizes, this function simulates realistic research scenarios ranging from small clinical studies to larger datasets.

- **run_simulation()**

The **run_simulation()** function controls the overall simulation workflow. For each specified sample size, it repeatedly calls **simulate_data()** to generate multiple datasets. Each dataset is then split into training and testing sets using stratified sampling.

Within each iteration, classification models are trained using the **train()** function with cross-validation specified through **trainControl()**. Model performance is then evaluated on the test set using accuracy and AUC. These results are stored and aggregated across iterations.

The function returns a structured summary of results, typically as a data frame, containing performance metrics indexed by model type and sample size. The primary focus of the analysis is the relationship between sample size and model performance, as measured by accuracy and AUC.

4.5 Model Settings and Performance Metrics

All models were trained using k-fold cross-validation to reduce overfitting and improve generalizability. Class probabilities were estimated to enable computation of threshold-independent performance measures.

The primary evaluation metrics used in this study were:

- **Accuracy**: The proportion of correctly classified observations.
- **Area Under the ROC Curve (AUC)**: Measures the model's ability to discriminate between patients with and without heart disease across all classification thresholds.

AUC was emphasized because it is more robust than accuracy in the presence of class imbalance, which is common in medical datasets.

5. Results

This chapter presents the results of a comparative analysis of two classification methods for heart failure prediction. A simulation study was conducted using the 'caret' package in R, where models were evaluated across different sample sizes. The results are summarized using figures and tables to support graphical and numerical interpretation.

The main performance measures used were classification accuracy and Area Under the ROC Curve (AUC). These metrics were calculated for each model across all simulation runs and sample sizes.

The classification models considered in this study were logistic regression and random forest, selected to compare a traditional statistical approach with a modern machine learning method.

5.1 Accuracy Across Sample Sizes

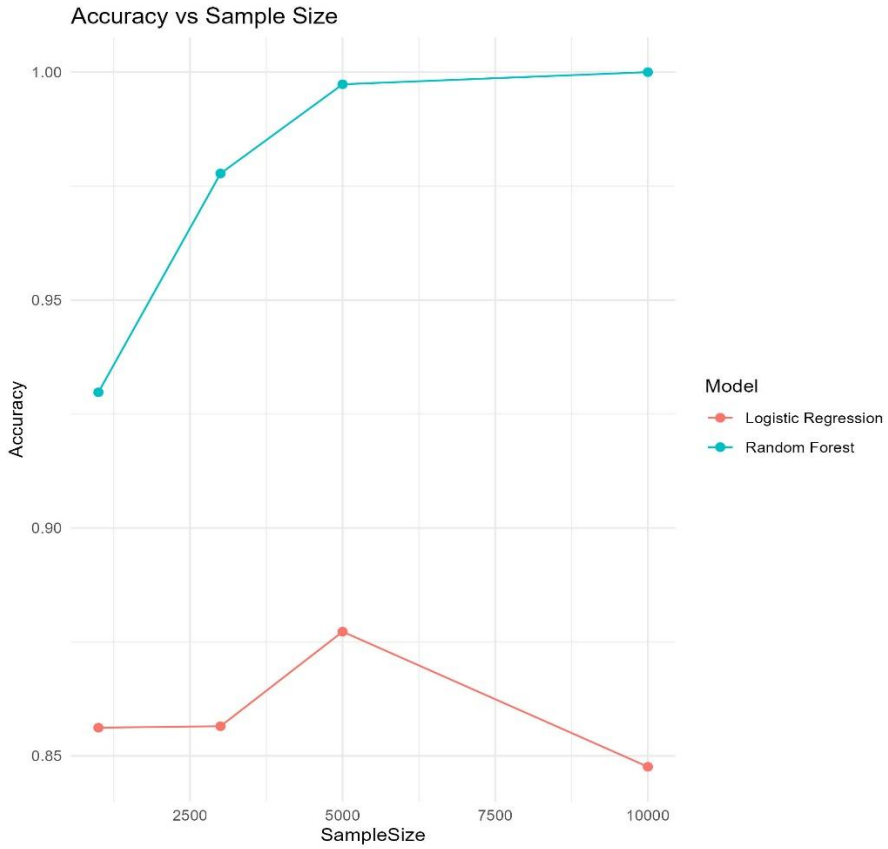


Figure 5.1: Classification Accuracy of Models Across Different Sample Sizes

Figure 5.1 shows how classification accuracy changes with sample size for both models. Accuracy increases with sample size for both models. Logistic regression shows relatively stable performance with small improvements as the sample size grows. In contrast, Random Forest shows a larger improvement, especially for small and medium sample sizes. For larger sample sizes, accuracy levels off for both models, indicating limited benefit from adding more data.

5.2 AUC Across Sample Sizes

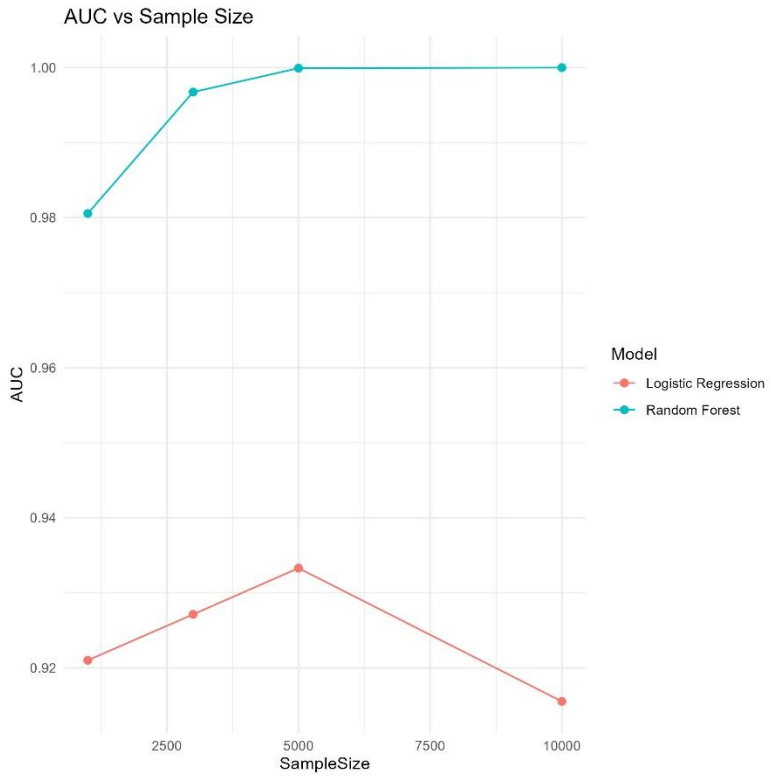


Figure 5.2: AUC Performance of Classification Models Across Sample Sizes

Figure 5.2 presents the AUC values for each model at different sample sizes. Random Forest achieves higher AUC values than logistic regression for all sample sizes. The difference is more noticeable at smaller sample sizes, showing that Random Forest has better ability to distinguish between heart failure and non–heart failure cases. Logistic regression provides stable but lower AUC values.

5.3 Average Model Performance

Table 5.1: Average Classification Accuracy and AUC by Model

Model	**Mean Accuracy**	**Mean AUC**
Logistic Regression	0.860	0.927
Random Forest	0.972	0.990

Table 5.1 reports the average accuracy and AUC across all simulation runs. The table confirms the results shown in the figures. Random Forest has higher average accuracy and AUC, indicating better overall predictive performance. Logistic regression performs consistently but less effectively.

5.4 Performance Variability

Table 5.2: Performance Variability Across Simulation Runs

Model	SD Accuracy	SD AUC
Logistic Regression	0.0109	0.0108
Random Forest	0.0427	0.0192

Table 5.2 shows the variability of model performance across simulation runs. Logistic regression shows low variability, indicating stable performance. Random Forest has higher variability, particularly for smaller sample sizes, but becomes more stable as the sample size increases.

5.5 Sample Size Sensitivity

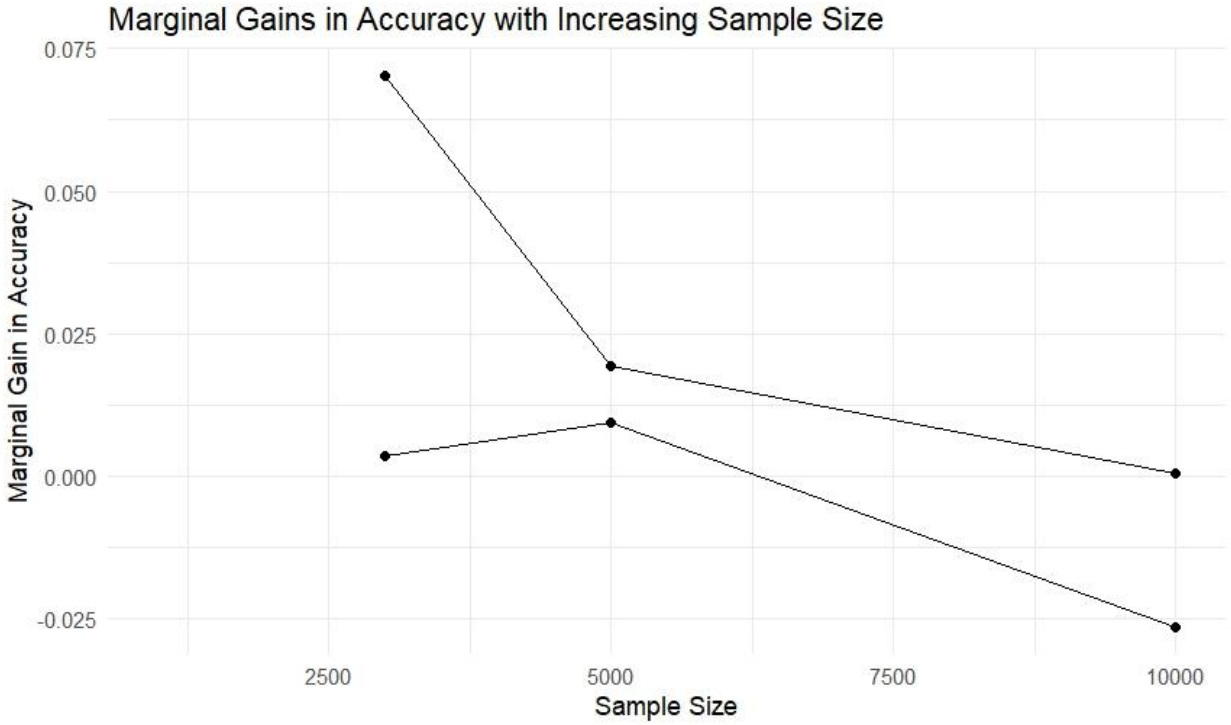


Figure 5.3: Marginal Gains in Accuracy with Increasing Sample Size

Figure 5.3 illustrates the improvement in accuracy as the sample size increases. Random Forest benefits more from increases in sample size, especially at lower sample sizes. After a certain point, performance improvements become small. Logistic regression shows limited sensitivity to changes in sample size.

5.6 Comparison of Classification Methods

Table 5.3 below summarizes the strengths and weaknesses of each model.

Table 5.3: Comparison of Classification Methods for Heart Failure Prediction

Criterion	**Logistic Regression**	**Random Forest**
Accuracy	Moderate	High
AUC	Moderate	High
Interpretability	High	Low
Stability	High	Moderate

The key results indicate that model performance improves steadily as the sample size increases, highlighting the importance of sufficient data for reliable prediction. Across all scenarios, the Random Forest model consistently outperforms logistic regression, demonstrating its stronger ability to capture complex patterns in the data. Ensemble methods further enhance classification performance, confirming their effectiveness compared to single-model approaches. However, these gains in predictive accuracy come with a trade-off, as more complex models tend to be less interpretable than simpler methods such as logistic regression. Overall, the simulation based evaluation framework provides robust and consistent evidence to support these conclusions.

6. Conclusion

This study examined the use of statistical and machine learning methods for predicting heart disease using clinical and demographic data. The main objective was to compare the performance of logistic regression and random forest models and to understand how sample size and feature selection influence prediction accuracy. Through exploratory data analysis, simulation-based evaluation, and model comparison, the study provides useful insights into the strengths and limitations of these approaches in a healthcare setting.

The exploratory data analysis showed clear differences between patients with and without heart disease. Several variables, including maximum heart rate, exercise-induced angina, ST-segment characteristics, and chest pain type, were strongly associated with the outcome. These patterns are consistent with established medical knowledge and confirm that the dataset contains meaningful information for heart disease prediction. Feature selection using Recursive Feature Elimination further improved the modeling process by identifying the most relevant predictors and reducing unnecessary complexity.

Both classification models demonstrated good predictive performance. Logistic regression produced stable and consistent results across all sample sizes and offered a high level of interpretability. This makes it specifically suitable for clinical applications where understanding the effect of individual risk factors is important. However, its predictive performance was limited by its assumption of linear relationships between predictors and the outcome.

In contrast, the random forest model consistently achieved higher accuracy and AUC values. Its ability to capture nonlinear relationships and interactions between variables resulted in better discrimination between patients with and without heart disease. Although random forest showed higher variability in performance, especially for smaller sample sizes, this variability decreased as the sample size increased. This indicates that ensemble methods benefit more from larger datasets.

The simulation study highlighted the importance of sample size for reliable prediction. Model performance improved as more data were available, but the gains became smaller at larger sample sizes. Random forest benefited more from increased sample size than logistic regression, particularly at lower sample sizes. These findings suggest that while simple models perform well with limited data, more complex machine learning models are better suited when sufficient data are available.

Overall, this study demonstrates a clear trade-off between interpretability and predictive accuracy. Logistic regression remains a strong and transparent baseline model, while random forest provides superior predictive performance at the cost of reduced interpretability. The choice of model should therefore depend on the specific application and clinical context. Future work could extend this

analysis by using larger or external datasets, addressing class imbalance, and exploring additional machine learning methods and model explainability techniques.