Statistical Theory: Heart Failure Prediction

Abstract

Heart disease is a leading cause of death worldwide, and accurate prediction models are essential for early diagnosis and prevention. This study aimed to develop and compare several machine learning models to predict the presence of heart disease using demographic and clinical data from 918 patients. We applied Logistic Regression, Random Forest, and Support Vector Machine (SVM) models, assessing their performance using accuracy, precision, recall, F1 score, and AUC-ROC metrics. While all models demonstrated good predictive ability, Random Forest slightly outperformed the others with an accuracy of 86% and the highest F1 score. Additionally, feature importance analysis highlighted ST slope, maximum heart rate, and age as key predictors of heart disease. A Friedman test revealed significant differences between the models' AUC-ROC scores, with SVM performing lower than the others. Despite strong results, limitations such as dataset size and potential biases from data preprocessing suggest that further validation and refinement are needed to improve the generalizability of these models.

Introduction

Cardiovascular diseases (CVDs) continue to be the leading cause of mortality worldwide, with approximately 17.9 million deaths annually (World Health Organization, 2021). Early detection and risk prediction are essential for implementing preventive measures and reducing mortality rates. Recent advancements in machine learning offer powerful tools for analyzing complex medical data and predicting health outcomes, including heart disease.

The objective of this study is to evaluate multiple machine learning models for predicting heart disease based on various patient characteristics and clinical measurements. By using a comprehensive dataset, we aim to identify which model performs best in terms of predictive accuracy and which features are the most influential in determining heart disease risk.

Background

The application of machine learning in healthcare, particularly for predicting cardiovascular diseases, has gained significant momentum. Several studies have explored different algorithms to predict heart disease with promising results. Despite the progress, there is a need for more comprehensive studies comparing multiple machine learning models using diverse patient data.

Research Objectives

1. To develop and evaluate multiple machine learning models (including logistic regression, random forest, and SVM) for predicting heart disease risk.

2. To identify the most significant predictors of heart disease from the available patient characteristics and clinical measurements, using methods such as logistic regression coefficients and feature importance from random forest.

3. To compare the performance of different models using rigorous statistical methods, including cross-validation and the Friedman test.

Methods

Dataset Description

The dataset used in this study consists of 918 patient records, each containing 12 features related to demographic information and clinical measurements. The target variable is binary, indicating the presence (1) or absence (0) of heart disease.

The features included in the dataset are:

1. Age (continuous)
2. Sex (categorical)
3. Chest Pain Type (categorical)
4. Resting Blood Pressure (continuous)
5. Serum Cholesterol (continuous)
6. Fasting Blood Sugar (binary: 1 = true, 0 = false)
7. Resting Electrocardiographic Results (categorical)
8. Maximum Heart Rate Achieved (continuous)
9. Exercise Induced Angina (binary: 1 = yes, 0 = no)
10. ST Depression Induced by Exercise Relative to Rest (continuous)
11. Slope of the Peak Exercise ST Segment (categorical)

Data Preprocessing

An initial examination of the dataset revealed no missing values. However, we identified zero values in two columns that are physiologically implausible:

- Cholesterol: 18.74% zero values
- Resting Blood Pressure (RestingBP): 0.11% zero values

To address this issue, we replaced these implausible zero values with the mean value of the respective column, excluding the zeros.

Exploratory Data Analysis (EDA)

Following the initial data cleaning, we conducted an exploratory data analysis to understand the characteristics and distributions of our variables.

1. Normality Test (Hypothesis Testing): We conducted the Shapiro-Wilk test to assess the normality of the continuous variables.

- Null hypothesis (H0): The data is normally distributed.
- Alternative hypothesis (H1): The data is not normally distributed.
- A significance level (α) of 0.05 was used.

2. Spearman's Correlation Analysis (Continuous Variables): After confirming that the continuous variables are not normally distributed, we used Spearman’s rank correlation to assess relationships between them.

3. Mann-Whitney U Test (Hypothesis Testing for Continuous Variables): To further assess whether continuous variables differ significantly between patients with and without heart disease, we conducted the Mann-Whitney U test, which compares the medians of two independent groups.

- Null hypothesis (H0): The distribution of the continuous variable is the same in both groups.
- Alternative hypothesis (H1): The distributions differ between the groups.
- A significance level (α) of 0.05 was used.

4. Chi-Square Test and Cramér's V (Categorical Variables): For categorical variables, we applied the Chi-Square test to examine their relationship with heart disease. The null hypothesis (H0) is that the categorical variable is independent of heart disease, while the alternative hypothesis (H1) is that there is an association between the variable and heart disease. A significance level (α) of 0.05 was used.

To measure the strength of these associations, we calculated Cramér's V, a measure of effect size for Chi-Square tests.

Model Building

To predict the presence of heart disease, we built and compared multiple classification models. The models included:

1. Logistic Regression: A basic logistic regression model was used to establish a baseline for comparison.
2. Logistic Regression with Interaction Terms: We explored the potential interactions between variables by adding interaction terms to the logistic regression model to assess whether these interactions improved predictive performance.
3. Random Forest: An ensemble model that uses multiple decision trees to improve prediction accuracy and reduce overfitting.
4. SVM: A classification model that seeks to find the hyperplane that best separates the classes.

For models that are sensitive to the scale of the input features, we applied feature scaling using the Standard Scaler from the scikit-learn library.

All models were evaluated using accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). We also computed confusion matrices to visualize the models' performance on classifying heart disease.

Cross-Validation: To improve the reliability of the model evaluations, we applied 5-fold cross-validation. This technique divides the dataset into five subsets (folds), trains the model on four folds, and tests it on the remaining fold, repeating the process five times. The average performance across all five folds is used to assess the model's overall accuracy and robustness.

Friedman Test (Comparison of Models): To compare the performance of the models we used the Friedman test, a non-parametric test designed to detect differences between dependent samples.

- Null hypothesis (H0): There is no significant difference in performance across the models.
- Alternative hypothesis (H1): At least one model performs significantly differently from the others.
- A significance level (α) of 0.05 was used.

Results

Continuous Variables Analysis

Normality Test: The results showed that all continuous variables have p-values lower than 0.05, leading to the rejection of the null hypothesis. This indicates that the data is not normally distributed, necessitating the use of non-parametric tests in further analyses.

Shapiro-Wilk test results:

Variable	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
p-value	2.165e-05	1.743e-12	1.311e-22	0.00017	8.272e-28

Spearman's Correlation Analysis: The analysis revealed several significant relationships. Notably, age showed a negative correlation with maximum heart rate (r = -0.37) and a positive correlation with ST depression (Oldpeak) (r = 0.30). Maximum heart rate demonstrated a negative correlation with heart disease (r = -0.40), while ST depression (Oldpeak) showed a positive correlation (r = 0.42). These relationships provide initial insights into potential factors influencing heart disease risk. (Figure 1)

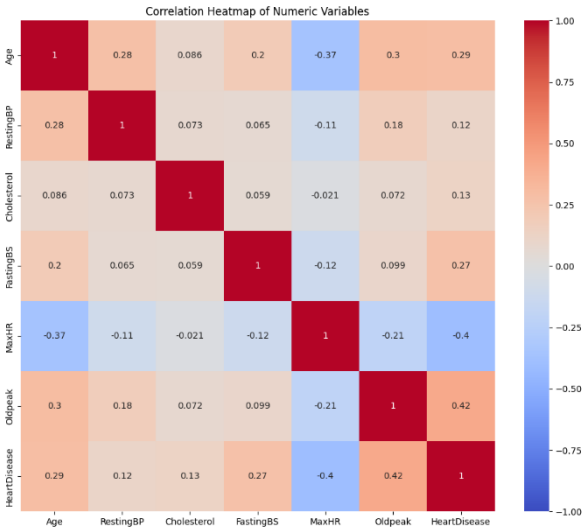

Figure 1: Spearman correlation heatmap

Mann-Whitney U Test: The test revealed statistically significant differences ($p < 0.05$) in all continuous variables between groups with and without heart disease. These findings suggest that all these variables could be potential predictors of heart disease, with particularly high significance observed for age, maximum heart rate (MaxHR), and Oldpeak.

Mann-Whitney U test results:

Variable	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
p-value	1.81e-18	0.000443	0.000110	1.51e-34	6.77e-37

Categorical Variables Analysis

Chi-Square Test Results: The test yielded significant results for all six categorical variables, as indicated by p-values below the significance level ($\alpha = 0.05$), suggesting that these variables are significantly associated with heart disease.

Variable	Sex	ChestPainType	FastingBS	RestingECG	ExerciseInducedAngina	ST Slope
Chi2	84.1451	268.0672	64.3207	10.9315	222.2594	355.9184
p-value	< 0.0001	< 0.0001	0.0001	0.0042	< 0.0001	< 0.0001

Cramér's V Results: To further assess the strength of the association between each categorical variable and heart disease, we calculated Cramér's V. The results indicate varying levels of association. (Figure 2)

Among the variables, ST Slope and Chest Pain Type show the strongest associations with heart disease, while Resting Electrocardiographic Results exhibits the weakest association.

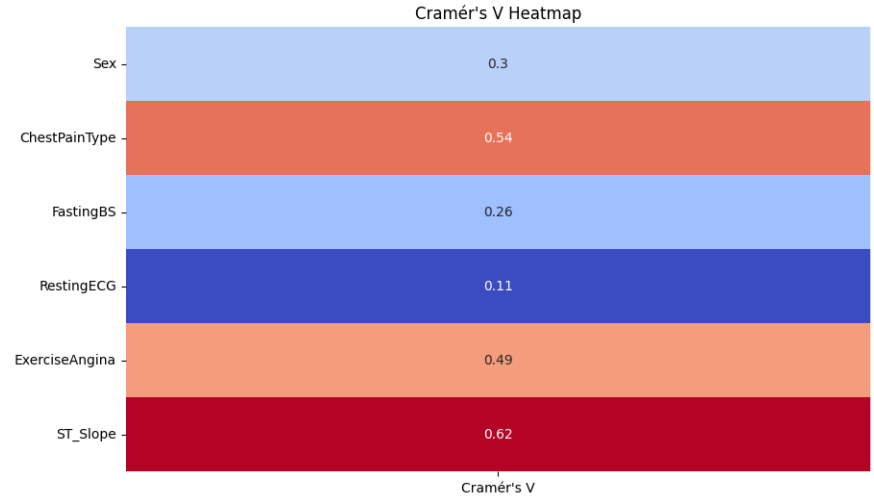

Figure 2: Heatmap of Cramér's V Values

Model Performance Comparison

We evaluated four classification models to predict the presence of heart disease: Logistic Regression, Logistic Regression with Interaction Terms, Random Forest, and Support Vector Machine (SVM). The performance metrics for each model, obtained through 5-fold cross-validation, are presented in the table below:

Comparison of Models:

	Metric	Logistic Regression	Logistic Regression (Interaction)	Random Forest	SVM
0	Accuracy	0.84	0.85	0.86	0.84
1	Precision	0.85	0.85	0.86	0.84
2	Recall	0.87	0.88	0.89	0.88
3	F1	0.86	0.86	0.87	0.86
4	AUC-ROC	0.92	0.92	0.92	0.90

As shown in the table, all models demonstrated strong performance, with accuracy ranging from 0.84 to 0.86. The Random Forest model slightly outperformed the other models across most metrics, achieving the highest accuracy (0.86), precision (0.86), recall (0.89), and F1 score (0.87).

To visualize the models' performance, we generated confusion matrices for each model (Figure 3) and compared their ROC curves (Figure 4).

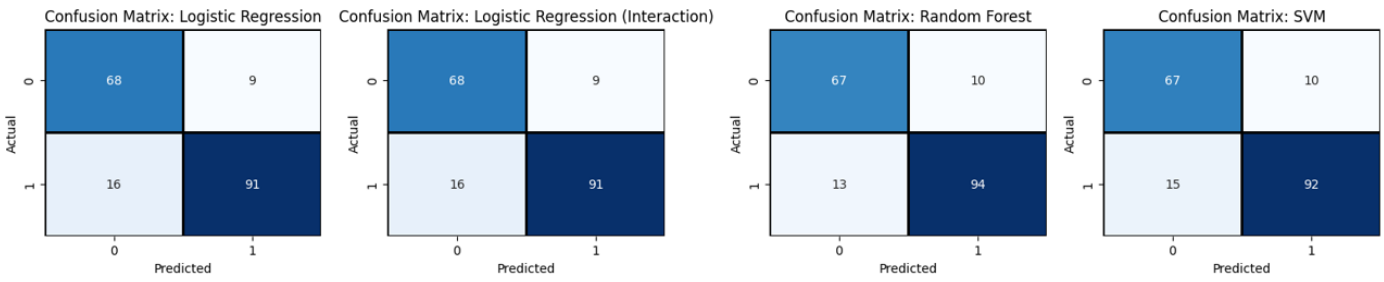

Figure 3: Confusion Matrices

Figure 3 illustrates that all models show similar patterns in their predictions, with the Random Forest model having slightly fewer misclassifications.

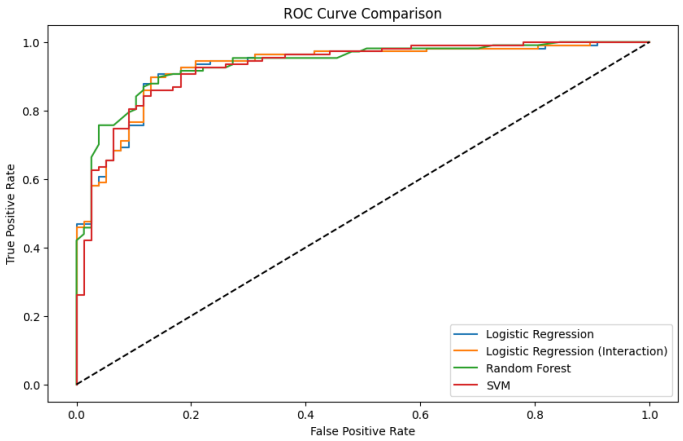

Figure 4: ROC curve comparison

Figure 4 demonstrates that all models have similar ROC curves, with high true positive rates achieved at relatively low false positive rates, confirming their strong predictive performance.

Friedman Test Results: To determine if the differences in model performance were statistically significant, we conducted a Friedman test for each performance metric. The results are as follows:

Metric	Accuracy	Precision	Recall	F1	Roc-auc
Statistic	2.5610	5.5909	0.7941	2.8636	9.7200
p-value	0.4644	0.1333	0.8509	0.4131	0.0211

The Friedman test results indicate that there is no statistically significant difference between the models in terms of accuracy, precision, recall, and F1 score ($p > 0.05$ for all). However, there is a significant difference in AUC-ROC scores ($p = 0.0211$), suggesting that at least one model performs differently from the others in terms of this metric.

Upon reviewing the AUC-ROC scores, the SVM model stands out with a lower score (0.90) compared to the other models (0.92). This indicates that the significant difference in the Friedman test is likely due to the SVM model's lower performance in this metric.

Feature Importance

To understand the factors contributing most to the prediction of heart disease, we examined the feature importances derived from the Random Forest model and the coefficients from the Logistic Regression models.

Figure 5 shows the feature importances from the Random Forest model. The top five most important features are ST_Slope_Up, Oldpeak, MaxHR, Age, and ST_Slope_Flat.

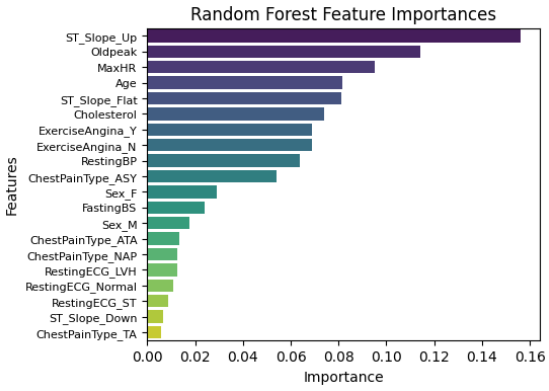

Figure 5: Random Forest Feature Importances

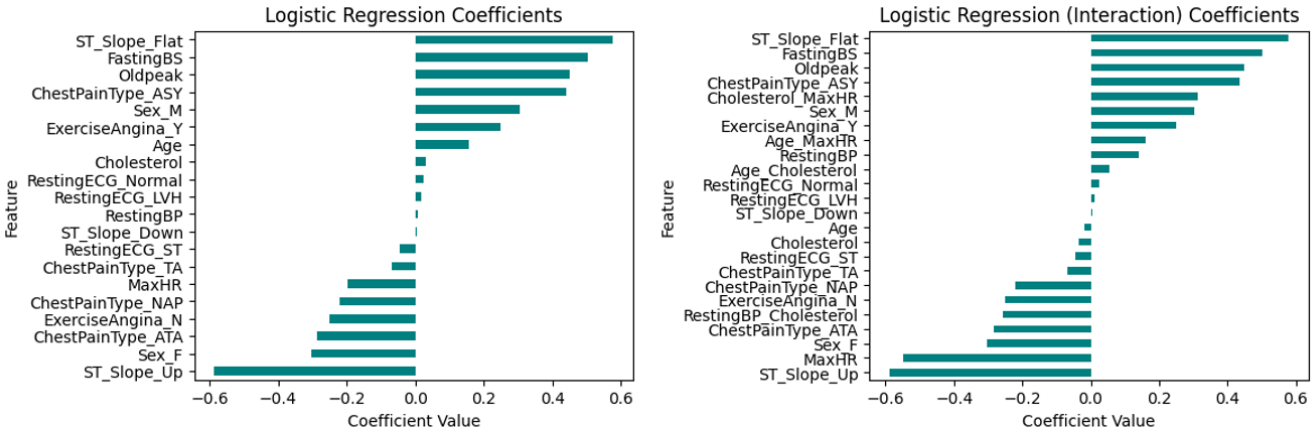

Figure 6: Logistic Regression Coefficients

Figure 6 displays the coefficients from both Logistic Regression models (with and without interaction terms). The most influential features, as indicated by the magnitude of their coefficients, are consistent with those identified by the Random Forest model.

These results suggest that ST segment characteristics, maximum heart rate, and age are among the most crucial factors in predicting the presence of heart disease.

Discussion

Conclusions

Model Performance: All four models (Logistic Regression, Logistic Regression with Interaction Terms, Random Forest, and SVM) demonstrated strong predictive performance, with accuracies ranging from 84% to 86%. The Random Forest model slightly outperformed the others across most metrics, although the Friedman test revealed that these differences were not statistically significant for accuracy, precision, recall, and F1 score. This suggests that all models are viable options for heart disease prediction, with Random Forest potentially offering a slight edge. The strong performance across all models underscores the potential of machine learning in supporting clinical decision-making for heart disease risk assessment.

Feature Importance and Correlations: Our analysis consistently identified several key factors as the most important predictors of heart disease across different methodologies:

- Model-based feature importance (Random Forest and Logistic Regression) highlighted ST segment characteristics (ST_Slope_Up, ST_Slope_Flat), ST depression induced by exercise (Oldpeak), maximum heart rate (MaxHR), and age as crucial predictors.
- Correlation analysis of continuous variables aligned with these findings, showing significant relationships between heart disease and age (positive), maximum heart rate (negative), and ST depression (positive).
- For categorical variables, Chi-Square tests and Cramér's V analysis revealed that all were significantly associated with heart disease, with ST Slope and Chest Pain Type showing particularly strong associations.

This consistency across different analytical approaches strengthens our confidence in these factors as key indicators of heart disease risk. The alignment between statistical correlations and model-derived importance suggests that these features are not only statistically significant but also valuable

in predictive modeling. These findings could guide healthcare providers in prioritizing specific indicators when assessing patient risk.

Limitations

1. Dataset Size and Representation: While our dataset included 918 patient records, a larger and more diverse dataset could potentially improve the generalizability of our findings. The demographic characteristics and geographic origin of the patients were not specified, which could limit the applicability of our models to different populations.

2. Data Quality: We encountered implausible zero values in the cholesterol and resting blood pressure columns, which we addressed by replacing them with mean values. However, this approach might have introduced some bias into our analysis. Future studies should aim to collect more accurate data or employ more sophisticated imputation techniques.

3. Model Complexity: While we explored several models, including those with interaction terms, there may be more complex relationships in the data that our models did not capture. Advanced techniques such as deep learning or more sophisticated ensemble methods could potentially uncover additional patterns.

4. Feature Set Limitations: Our analysis was limited to the existing features in the dataset. While we explored interactions, we did not create new features or include additional variables that might improve heart disease prediction. Important predictors, such as lifestyle factors or detailed medical history, may have been overlooked. Future studies should expand the feature set to enhance prediction accuracy.

5. Temporal Aspects: The dataset does not include temporal information, which could be valuable for understanding disease progression and improving prediction accuracy. Longitudinal studies could provide more comprehensive insights into heart disease risk factors and their evolution over time.

6. External Validation: Our study used cross-validation to assess model performance, but external validation on a completely independent dataset would provide a more robust evaluation of the models' generalizability.

In conclusion, our study demonstrates the potential of machine learning models in predicting heart disease risk and identifies key factors associated with heart disease. These findings could aid in early detection and risk stratification, potentially improving patient outcomes. While our models show promising results, the limitations highlighted suggest areas for future research, including larger and more diverse datasets, more advanced modeling techniques, and external validation studies. As machine learning continues to evolve, its integration with clinical expertise will be crucial in leveraging these tools effectively in healthcare settings.

Code and Analysis Replication

To reproduce our results, please follow these steps:

1. Access our shared Google Drive folder: Click here to access the project folder
2. Open the Colab notebook within the folder
3. Run all cells in the notebook sequentially