Identification of Risk Factors for Heart Disease Using Statistical Methods and Machine Learning

A20521362

Young Sun Lee

Abstract

Heart disease is one of the leading causes of death worldwide, and early prediction and prevention of its onset are crucial. With the advancement of modern medicine and data science, various medical records of patients are being digitized and stored. Analyzing such data to identify risk factors for heart disease can play a vital role in its prevention or early prediction. In this study, we analyzed the main factors of heart disease using the Kaggle heart disease dataset. Correlation coefficients were calculated for continuous variables, and chi-squared tests and Fisher's exact tests were conducted for categorical variables. Furthermore, continuous variables were categorized into 'high', 'medium', and 'low' and then transformed into categorical variables to identify key factors through chi-squared testing. Subsequently, various machine learning models, including logistic regression, decision trees, k-NN, random forests, gradient boosting, and SVM, were applied to build predictive models for heart disease and evaluate the performance of each model. The importance of variables for each model was analyzed, ranked, and visualized, revealing ST_Slope and Chest Pain as the most significant predictive factors.

1. Objective

The aim of this project is to analyze the characteristics of individuals with and without heart disease using statistical methods and machine learning models, and to determine the most significant factors in the onset of heart disease. To achieve this, methods such as correlation coefficients, chi-squared tests, and Fisher's exact tests will be employed. Additionally, machine learning models like logistic regression, decision trees, k-NN, random forests, gradient boosting, and SVM will be applied for evaluating model performance and assessing the importance of variables.

1.1 Specific Questions

What are the most important predictive variables for predicting heart disease?

How will the ranking be determined when both continuous and categorical variables are present?

What is the variable importance ranking among various prediction algorithms?

Among these algorithms, which one performs the best in predicting the onset of heart disease?

2. Dataset

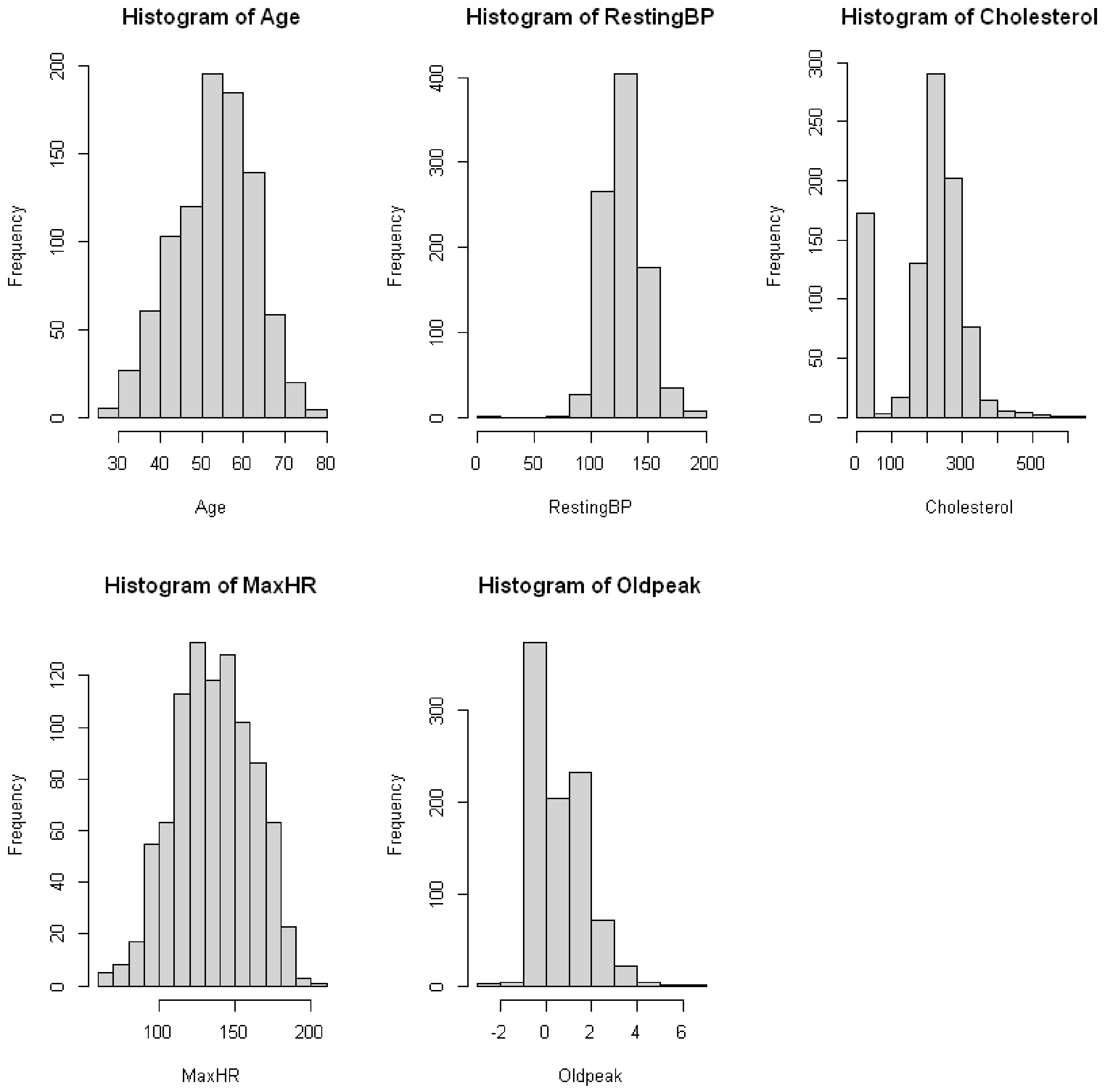
Heart Failure Prediction Dataset from Kaggle (originally from University of California machine learning repository). The dataset consist of 12 variables and 918 observations. The description of the variables are :

- **Age:** age of the patient [years]
- **Sex:** sex of the patient
    - M: Male
    - F: Female
- **ChestPainType:** chest pain type
    - TA: Typical Angina
    - ATA: Atypical Angina
    - NAP: Non-Anginal Pain,
    - ASY: Asymptonic
- **RestingBP:** resting blood pressure [mmHg]
- **Cholesterol:** serum cholesterol [mm/dl]
- **FastingBS:** fasting blood sugar
    - 1: if FastingBS > 120 mg/dl
    - 0: otherwise
- **RestingECG:** resting electrocardiogram results
    - Normal: Normal
    - ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05)
- **MaxHR:** maximum heart rate achived [Numeric value between 60 and 202]
- **ExerciseAngina:** exercise-induced angina

  - Y: Yes
  - N: No
- **Oldpeak:** oldpeak [Numeric value measured in depression]
- **ST_Slope:** the slope of the peak exercise ST segment
  - Up: upsloping
  - Flat: flat
  - Down: downsloping
- **HeartDisease:** output class
  - 1: heart disease
  - 0: normal

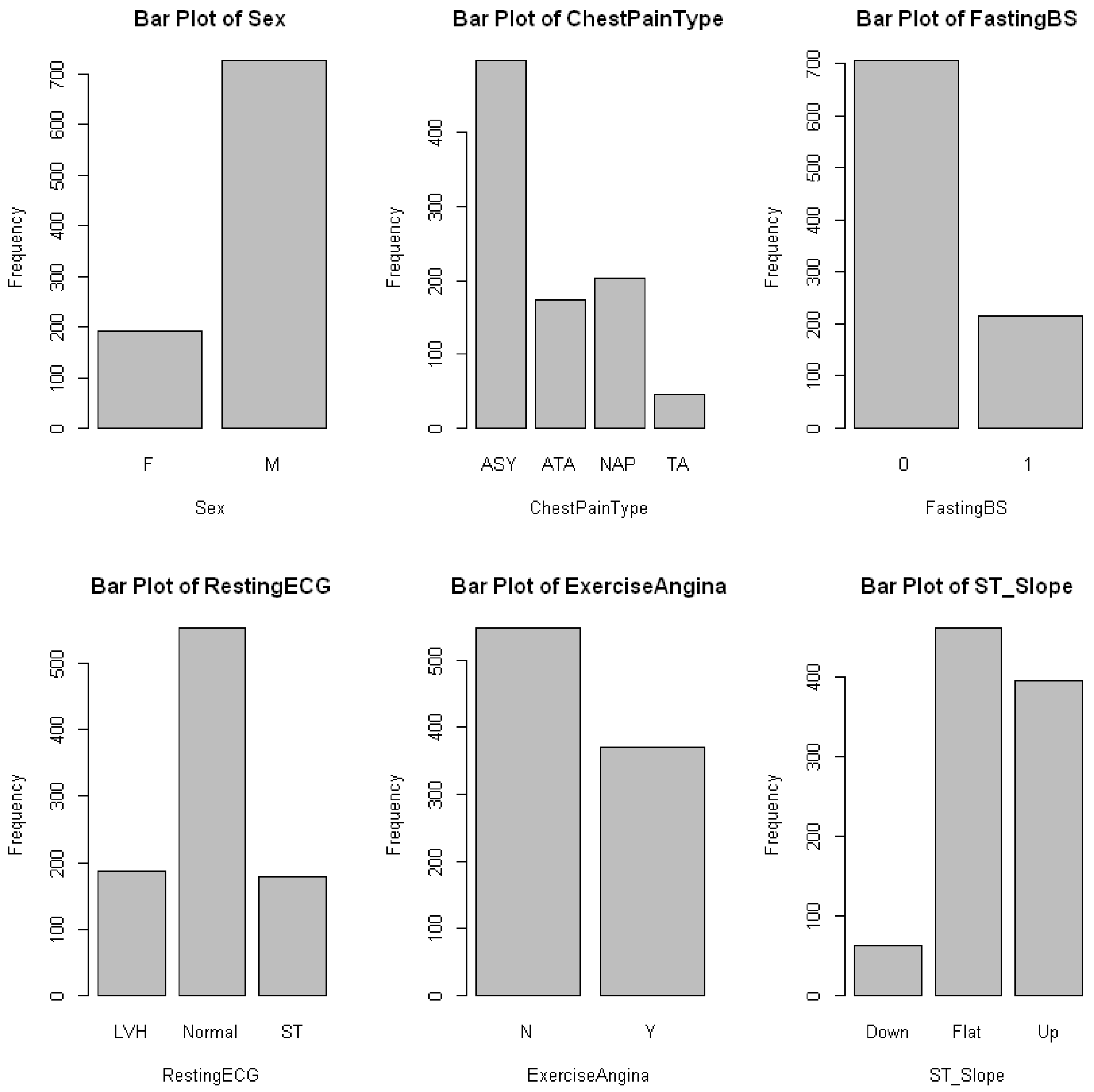
2.1 Distribution of continuous variables

There are outliers in Cholesterol.



2.2 Distribution of categorical variables

There are more men in data, and overall, the elements are not evenly distributed.



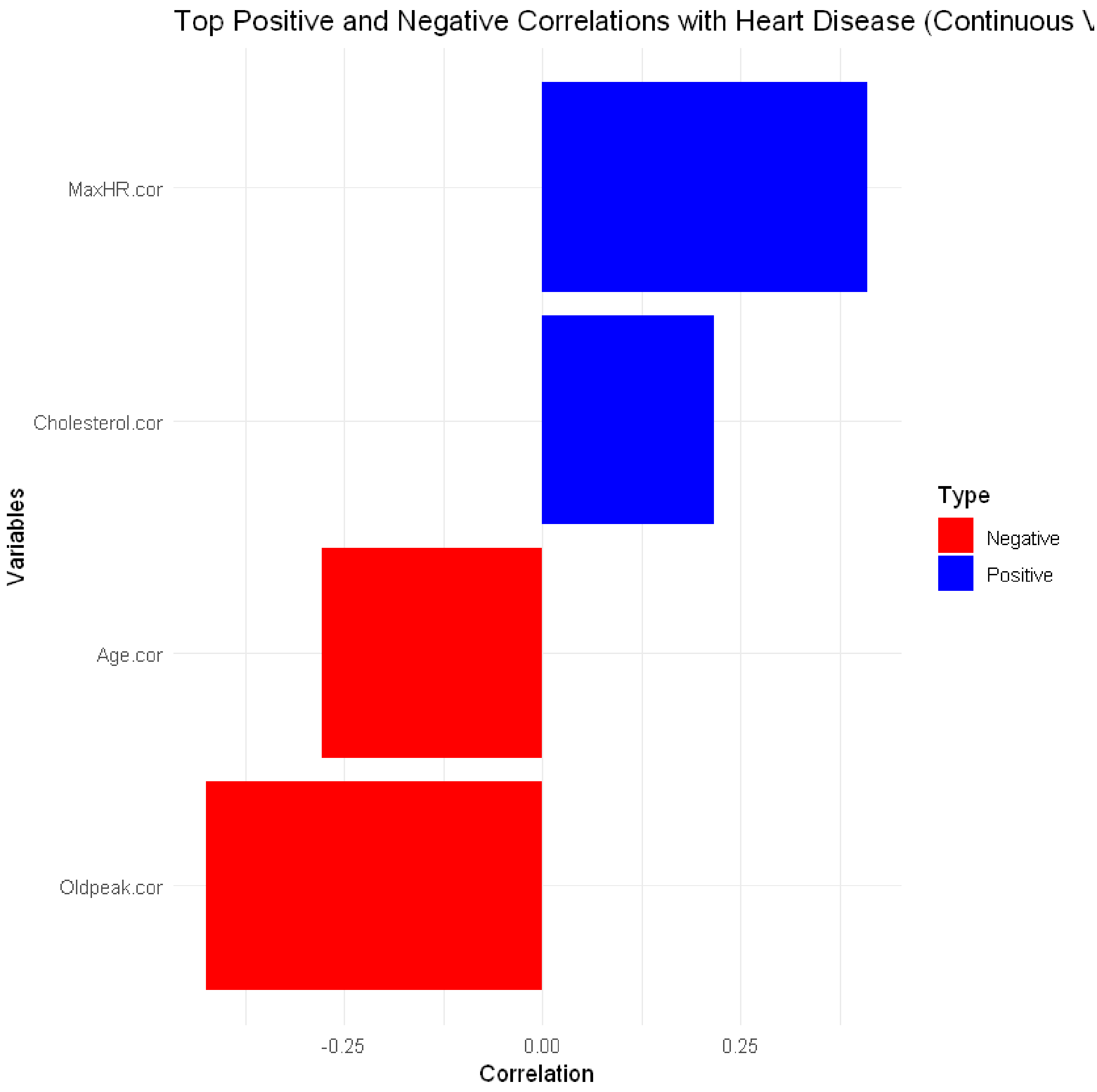
Subsequently, the categorical variables were converted to dummy variables.

The data was split into 80% training data and 20% test data.

3. Correlation

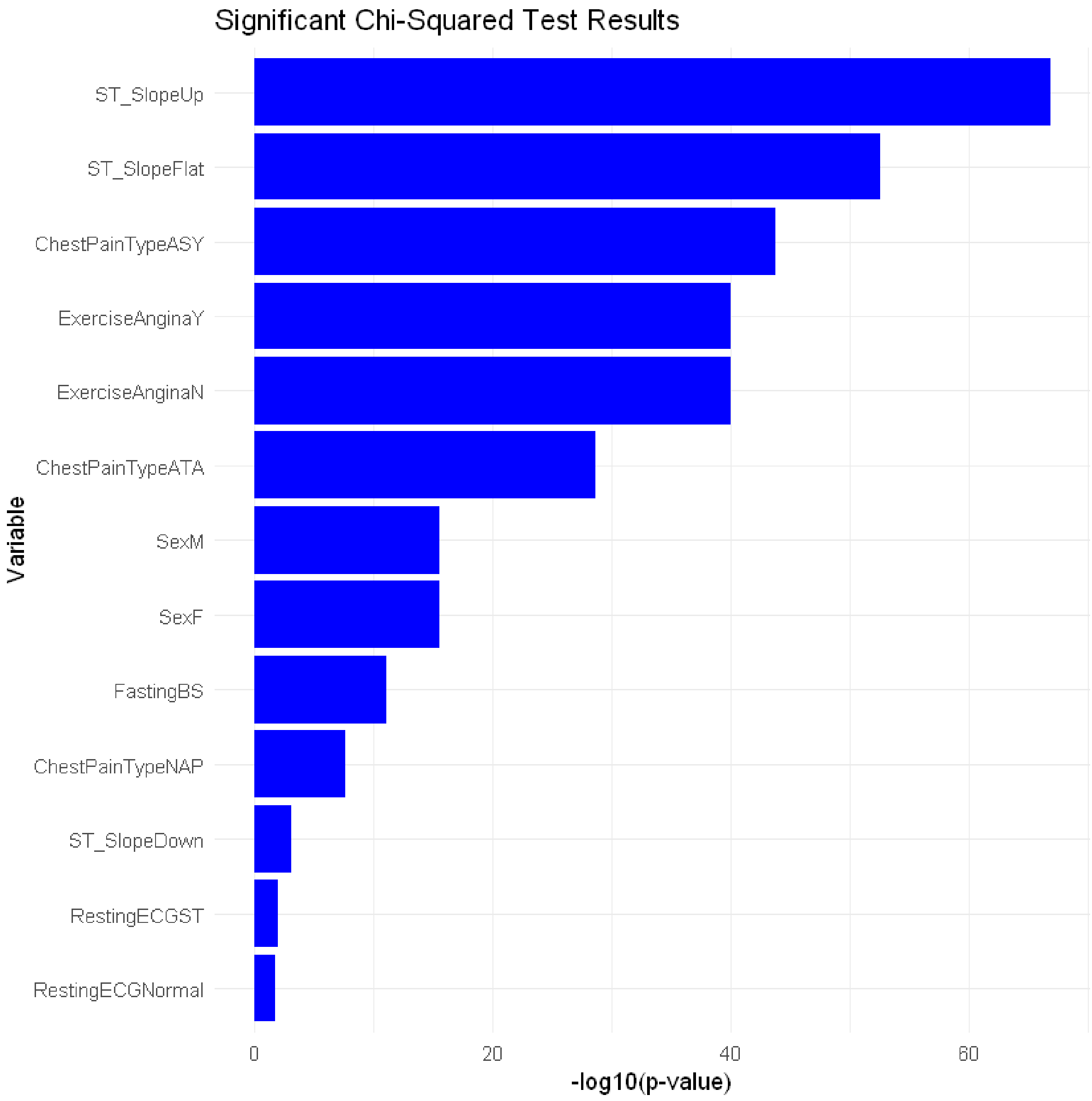
3.1 Correlation coefficient for continuous variables

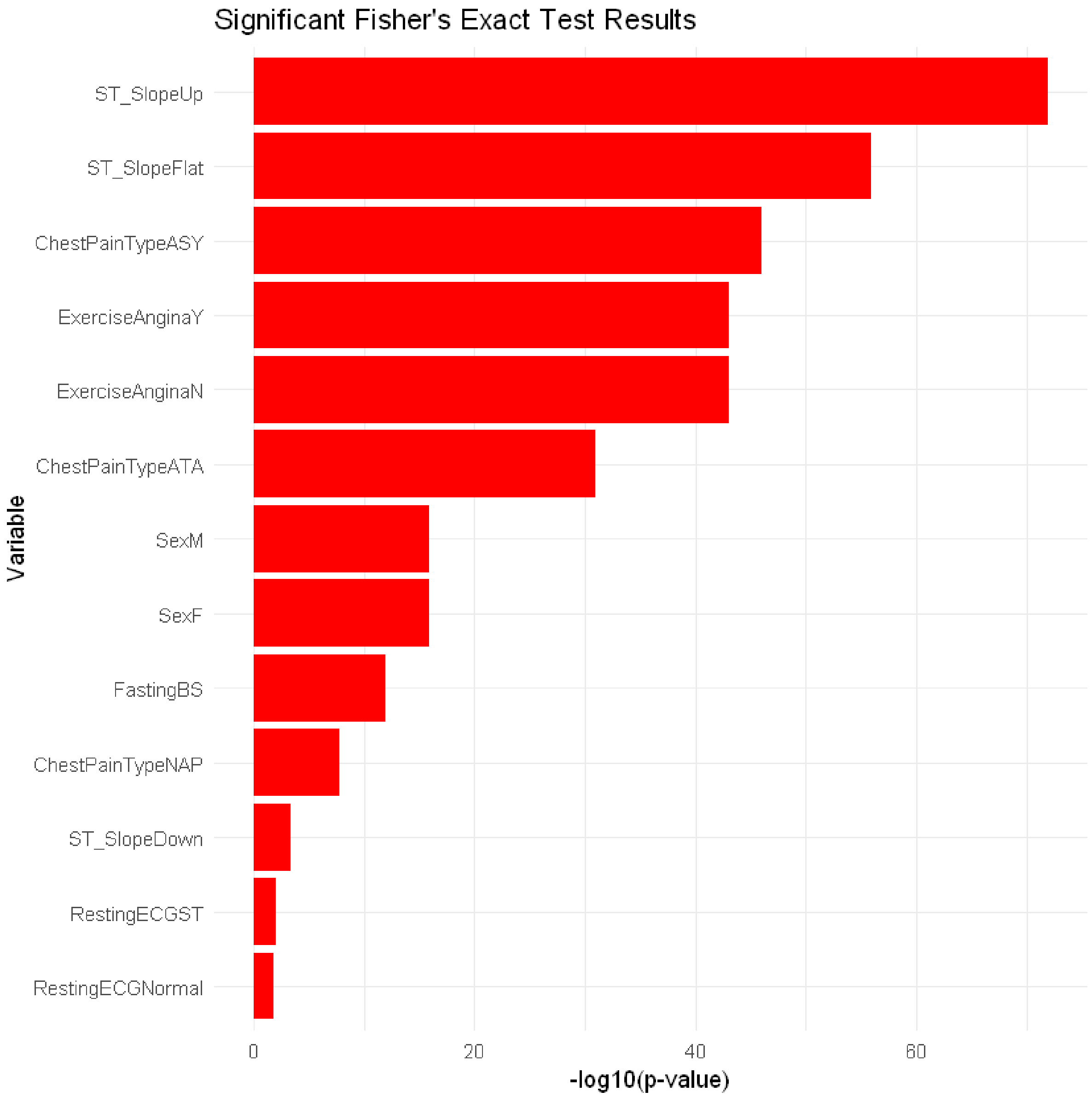
For continuous variables such as 'Age', 'RestingBP', 'Cholesterol', 'MaxHR', and 'Oldpeak', we calculated the point-biserial correlation coefficient with 'HeartDisease' (presence or absence of heart disease). This is a method for measuring the correlation between continuous variables and a binary categorical variable. Among the calculated correlation coefficients, those with an absolute value of 0.2 or higher were selected. The results showed that Maximum Heart Rate was the most significant positive factor, while Oldpeak was the most significant negative factor.



3.2 Chi-squared and Fisher’s exact tests for categorical variables

I generated frequency tables for categorical variables such as 'SexF', 'SexM', 'ChestPainTypeASY', etc., and conducted Chi-square tests and Fisher's exact tests. The Chi-square test and Fisher's exact test are statistical methods used to test the independence between two variables in categorical data. The null hypothesis assumes that variables A and B are independent, and if the Chi-square value is large, the null hypothesis is rejected.

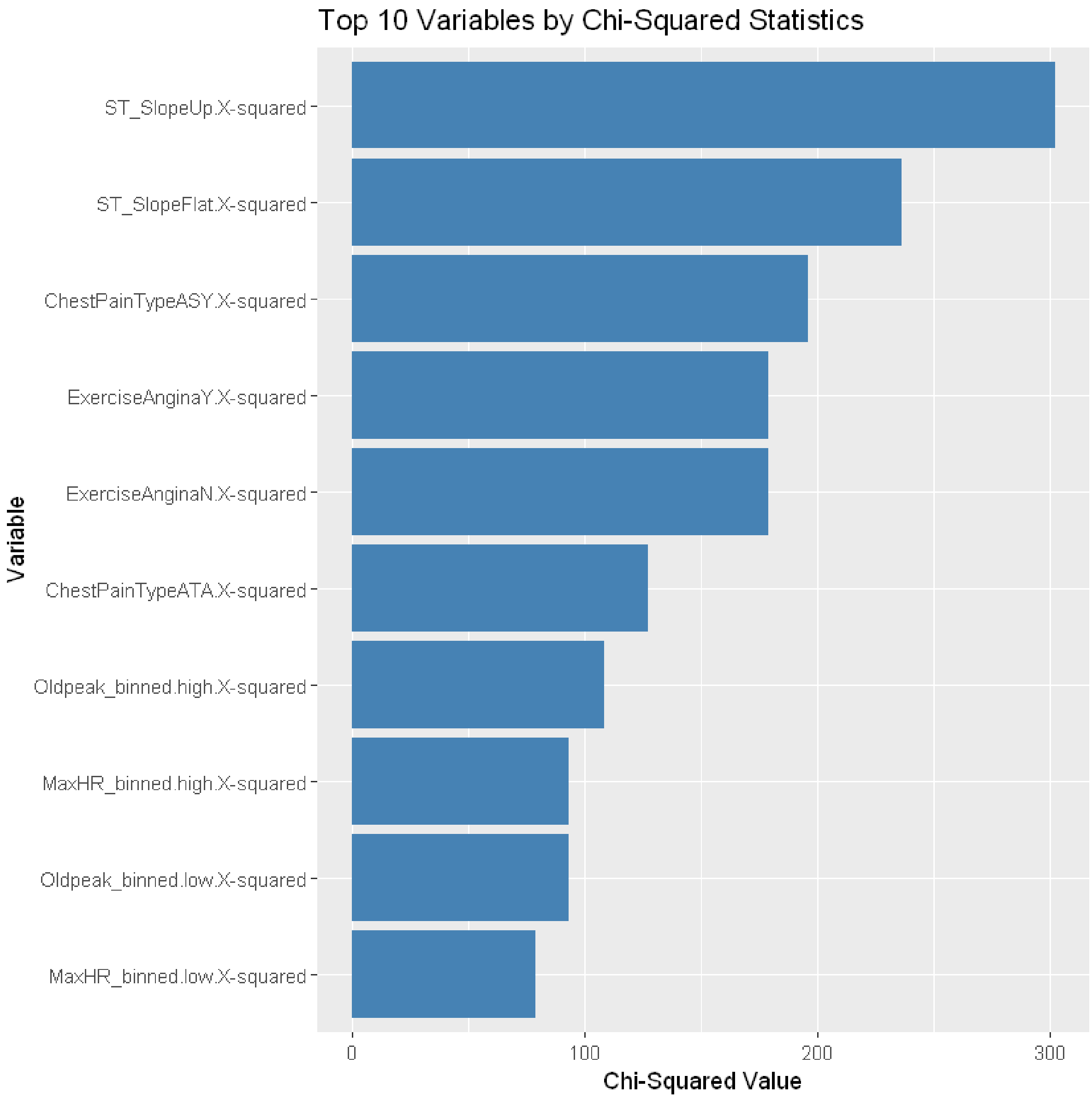




Both results show similar findings. ST_Slope demonstrates the highest association, followed by Chest Pain and Exercise Angina, which also show high correlations.

3.3 Chi-square Test with All Variables Categorized

To assess the rank order of correlations by integrating continuous and categorical variables, the continuous variables were categorized into high, medium, and low. After converting all into categorical variables, important variables were investigated using the Chi-square test.



In the rank order that combined the categorized continuous variables and categorical variables, the rankings of ST_Slope, Chest Pain, and Exercise Angina remained unchanged. It is noticeable that Maximum Heart Rate appeared somewhat lower in the ranking.

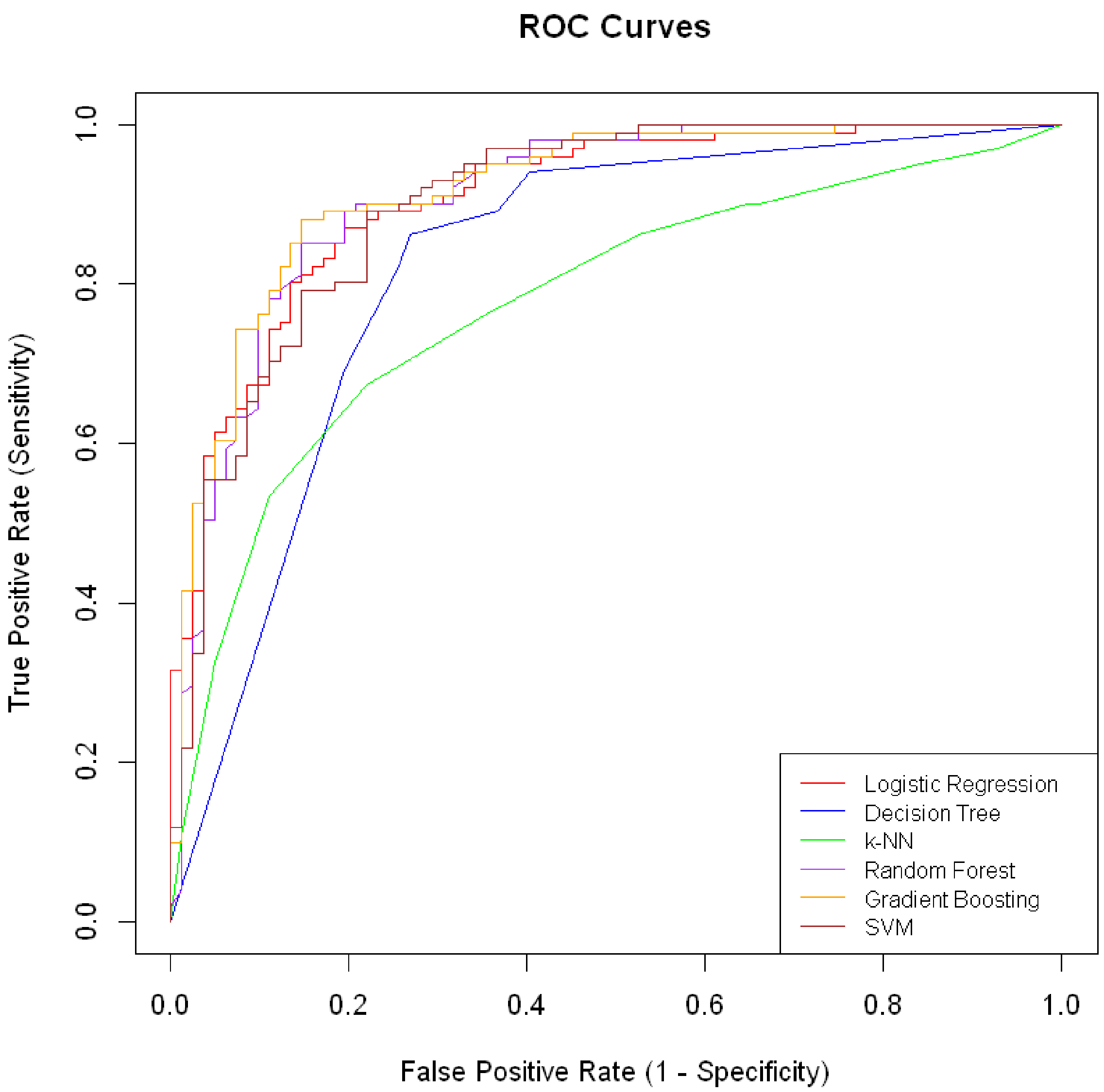
4. Model Performance Evaluation and Ranking of Variable Importance by Model

I build various machine learning models including Logistic Regression, Decision Tree, k-NN (k-Nearest Neighbors), Random Forest, Gradient Boosting, and SVM (Support Vector Machine). All models use the training data (trainData) to predict heart disease.

I evaluate the performance of the models using 5-fold cross-validation. For this purpose, we set up a 'trainControl' object.

The performance is assessed using the ROC curve, Performance Metrics, and AUC (Area Under the Curve).

4.1 ROC Curves



4.2 Performance Metrics

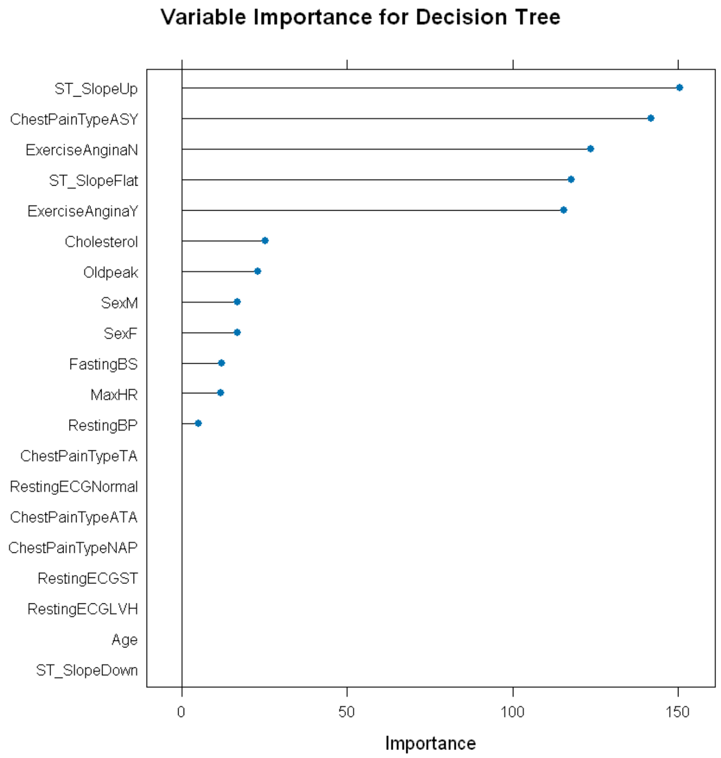
```
                    Precision    Recall  F1 Score
Gradient Boosting   0.8653846 0.8910891 0.8780488
Random Forest       0.8490566 0.8910891 0.8695652
Logistic Regression 0.8461538 0.8712871 0.8585366
SVM                 0.8285714 0.8613861 0.8446602
Decision Tree       0.7981651 0.8613861 0.8285714
k-NN                0.7264151 0.7623762 0.7439614
```

4.3 Area Under Curve

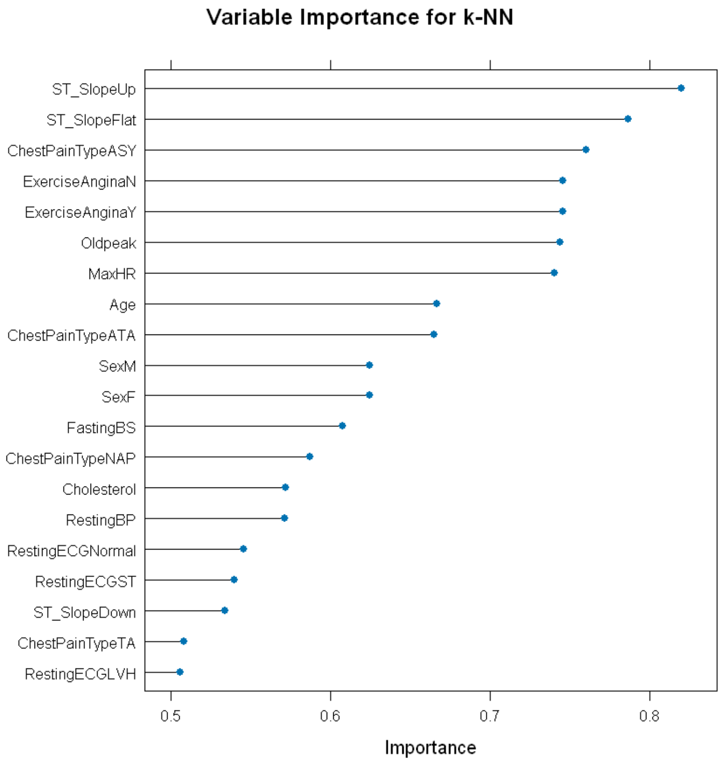
```
                Model       AUC
1   Gradient Boosting 0.9184979
2       Random Forest 0.9092007
3 Logistic Regression 0.9078725
4                 SVM 0.9020768
5       Decision Tree 0.8228689
6                k-NN 0.7779522
```

4.4 Variable importance by Model

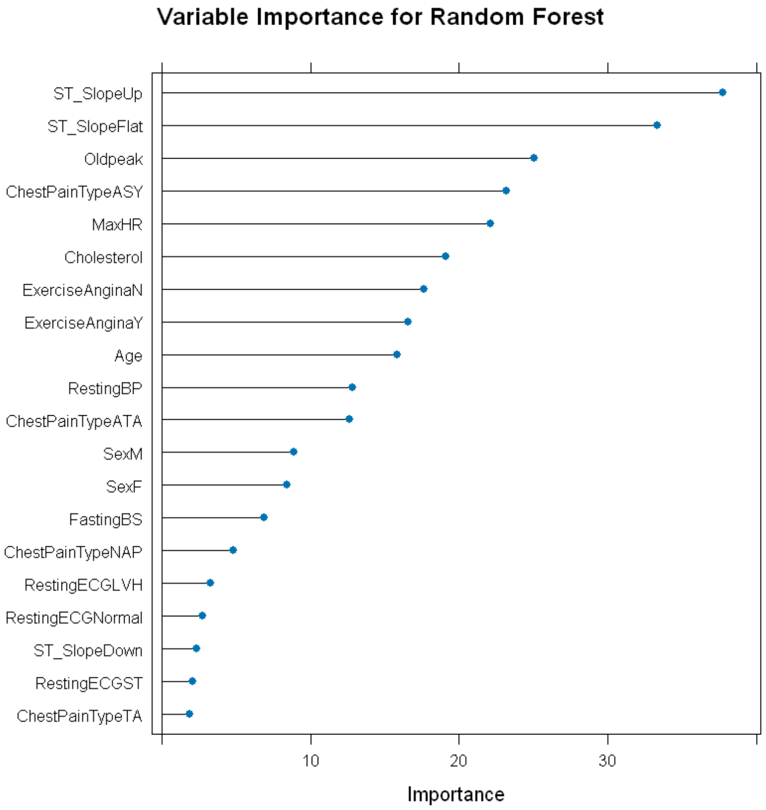
4.4.1 Decision Tree



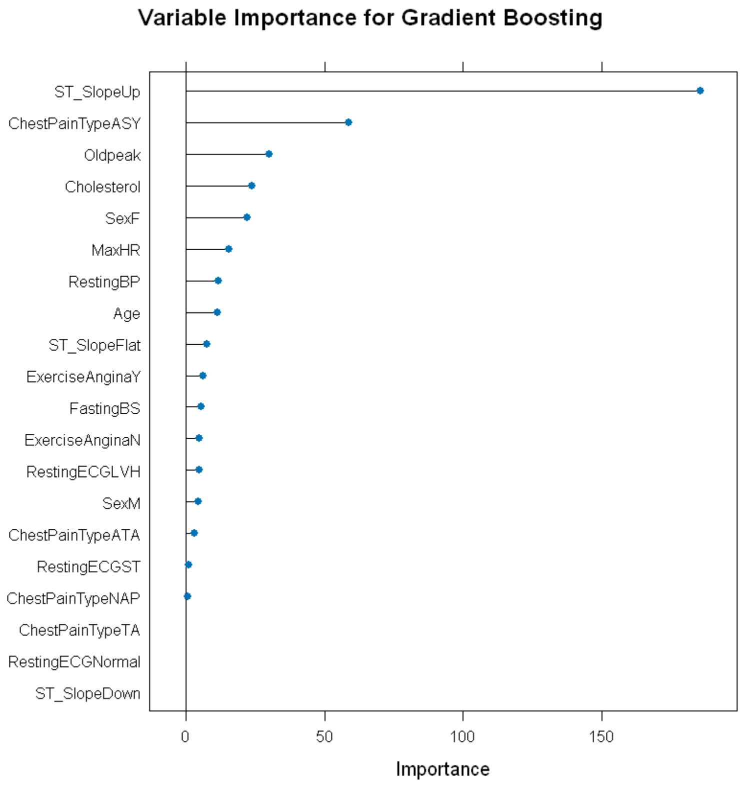
4.4.2 k-NN



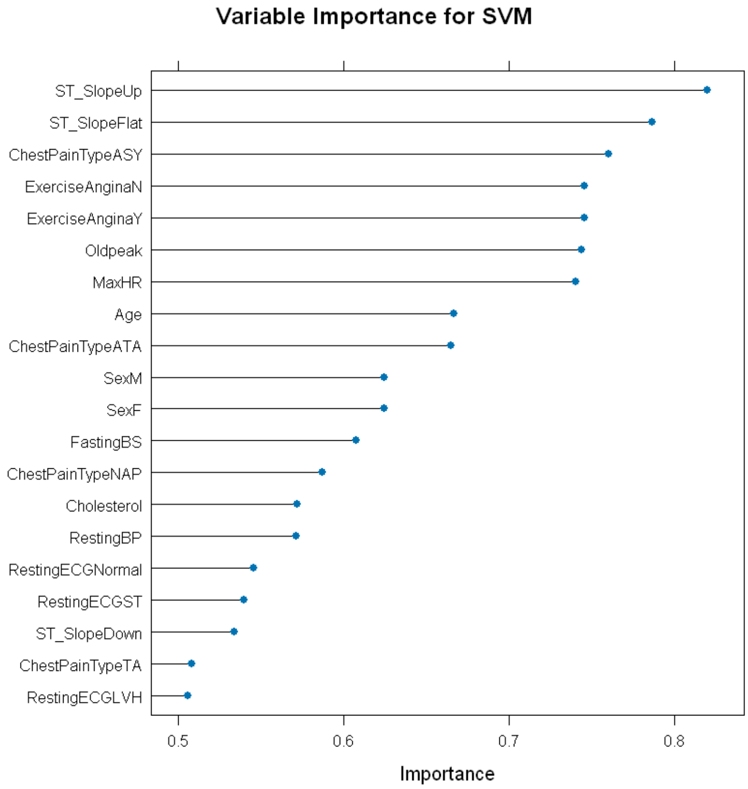
4.4.3 Random Rorest



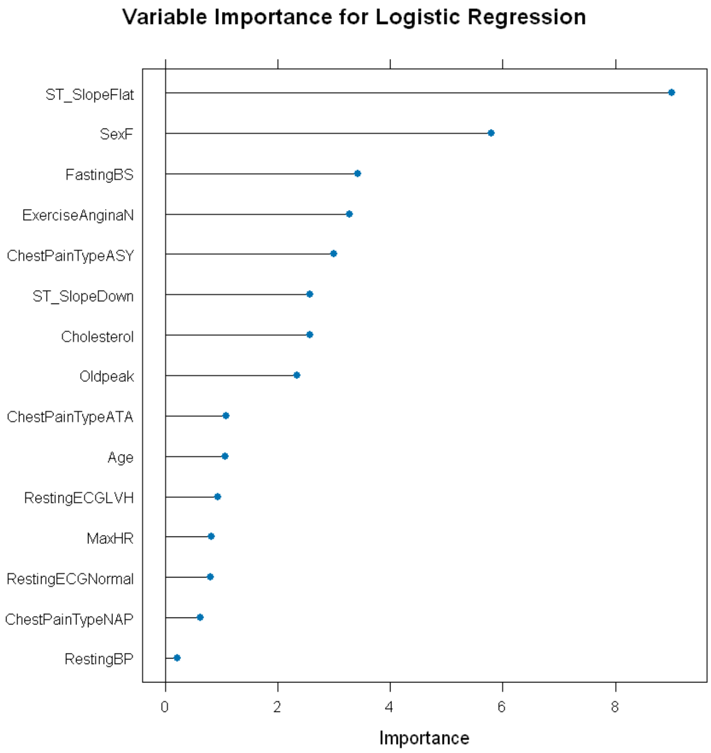
4.4.4 Gradient Boosting



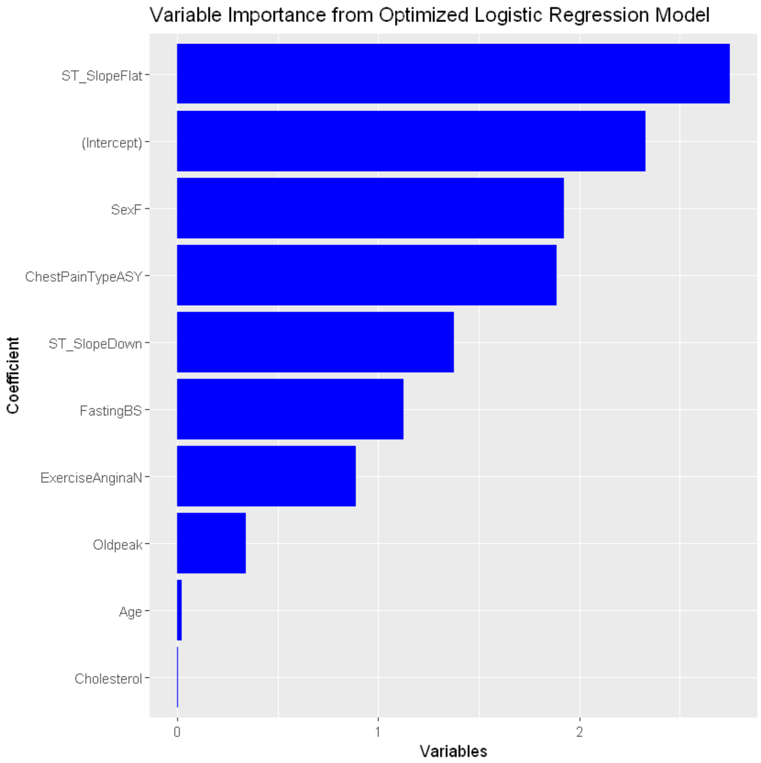
4.4.5 SVM



4.4.6 Logistic Regression

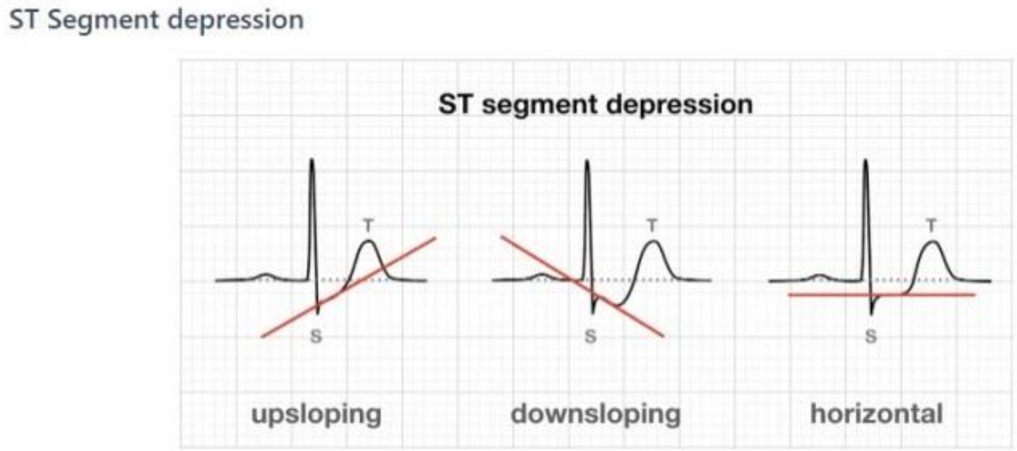


For Logistic Regression, I attempted optimization through stepwise regression. As a result, we were able to obtain the following outcomes.



In all models except Logistic Regression, ST_SlopeUp was identified as the most important factor. In Logistic Regression, the most important factor was ST_SlopeFlat, which is also in the ST_Slope category.

ST_Slope refers to a pattern observed in electrocardiograms (ECGs).



Upsloping ST-segment depression during physical exercise is usually normal, as long as the T-waves are not inverted. Hyperventilation can also cause similar ST-segment depressions. Conversely, Downsloping and Flat (horizontal) ST-segments typically indicate ischemia.

Thus, ECGs are a very strong predictor. According to the data from this project, for Downsloping, 49 out of 63 (77.8%) had heart disease, and for Flat, 381 out of 460 (82.8%) had heart disease. In contrast, only 78 out of 395 (19.7%) with Upsloping had heart disease.

The second most important factor in Decision Tree, K-NN, Gradient Boosting, and SVM was asymptomatic chest pain. This is also a strong indicator, as 392 out of 496 (79%) with symptoms had heart disease.

In Logistic Regression, the second most important factor was sexF, indicating a higher prevalence in women. In Random Forest, the second factor was Oldpeak, a numeric value measured in depression, also related to ECGs. Exercise Angina also consistently ranked high.

This aligns with the results of the Chi-square test and Fisher’s exact test on categorical variables.

5. Conclusion and discussion

In health check-ups, we can easily measure blood pressure, cholesterol levels, and blood sugar, and these are often considered strong predictors of heart disease. However, in this project, actual strong predictors of heart disease have been identified as exercise stress ECGs and chest pain. While blood pressure, cholesterol, and blood sugar may be related to heart disease in the long term, they are not the most accurate predictors. Therefore, regular ECGs might be the most important measure to prevent heart disease.

To conduct an exercise stress test, it typically requires about 15 minutes of activity on a treadmill or a bicycle ergometer. This is a significant hurdle for frequent testing in a hospital setting. Therefore, the development of a device that allows for easy monitoring of the electrocardiogram (ECG) during regular exercise, along with an application to interpret the data, could enable self-diagnosis. This approach is expected to significantly reduce the risk of heart diseases.

Interpreting ECGs is not a simple task. Abnormalities in ECGs can appear unexpectedly or be so subtle as to make judgment difficult. In this project, predictions based solely on ST_Slope stayed around 80%. This is why the U.S. Preventive Services Task Force (USPSTF) points out the limitations of ECG testing in screening for heart disease. However, recent studies using artificial intelligence to analyze ECG data have been published, showing potential in early detection of heart diseases. These studies could lead to more accurate predictions of heart diseases. For this, research with larger samples is needed, and efforts must be made to develop better algorithms.

Code in GitHub:

https://github.com/sun2423/CSP571_Project/

References

- Karna Vishnu Vardhana Reddy et al, Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators, Applied Sciences, 2021
- Mamun Ali et al., A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients, Healthcare Analytics, 2023.