## Analysis of Heart Disease Risk Factors Using Statistical Techniques and Predictive Modeling in R

## Objectives

- To explore the relationship between health parameters and the likelihood of heart disease.

- To compare logistic regression, decision trees, and random forests for heart disease prediction.

- To identify significant predictors and assess the performance of models.

## Introduction

This study investigates the influence of key factors such as age, cholesterol, chest pain type, and gender on heart disease. By combining statistical hypothesis testing and machine learning methods, the project evaluates predictive models and identifies crucial predictors.

## Dataset Description

Total observations: 303

Variables analyzed: Age, cholesterol (chol), chest pain type (cp), resting blood pressure (trestbps), maximum heart rate (thalach), gender (sex), and target (presence/absence of heart disease).

Data was split into training (80%) and testing (20%) sets.

## Statistical Analysis

### 1. T-Tests:

Cholesterol vs. Target: p-value = 0.1366, indicating no significant difference.

Age vs. Target: p-value = 7.061e-05, indicating a significant age difference between groups.

### 2. Chi-Square Tests:

Chest Pain vs. Target: p-value < 2.2e-16, strong association with heart disease.

Gender vs. Target: p-value = 2.667e-06, significant association with heart disease risk.

[Insert plot showing T-Test results]

## Predictive Modeling

### 1. Logistic Regression:

Formula: target ~ age + chol + cp + trestbps + thalach + sex.

Significant predictors: Chest pain (cp), age, and maximum heart rate (thalach).

Performance: Moderate accuracy observed in classification.

### 2. Decision Trees:

Method 1 (Tree):

- Error rate: 21.4%

- Accuracy: 78.6%

- Key predictors: Chest pain (100% importance) and age (51.44%).

Method 2 (Party):

- Constructed using ctree with predictors age, chol, cp, thalach, and exang.

- Clear visual splits based on variables of significance.

Method 3 (C50):

- Built using C5.0, providing interpretable splits.

- Accuracy and confusion matrix evaluated to determine performance.

### 3. Random Forest:

Formula: target ~ age + chol + cp + trestbps + thalach + sex.

- Optimal mtry: 2 (best accuracy from tuning).

- Accuracy: 73.35%.

- Kappa: 0.47 (moderate agreement).

- Key predictors: Chest pain (most significant), age, and maximum heart rate.

### Support Vector Machine (SVM):

Formula: target ~ age + chol + cp + trestbps + thalach + sex.

- Moderate performance in classifying individuals into target categories.

## Visualizations

Visualizations created in Power BI include:

- Scatterplots to explore relationships between variables.

- Bar charts for frequency distributions.

- Heatmaps to depict associations between categorical variables.

- Other standard visuals highlighting predictor importance.

## Findings

Key Results from Statistical Tests:

- No significant difference in cholesterol levels between groups (p-value = 0.1366).

- Significant differences in age between groups (p-value = 7.061e-05).

- Strong associations between chest pain and target (p-value < 2.2e-16) and gender and target (p-value = 2.667e-06).

Model Performances:

- Logistic regression: Baseline accuracy with interpretable coefficients.

- Decision trees: High interpretability, moderate accuracy (78.6%).

- Random forest: Improved accuracy (73.35%) and variable importance assessment.

- SVM: Moderately effective in prediction.

Variable Importance:

- Chest pain and age are the most significant predictors.

- Maximum heart rate is also influential in heart disease prediction.

## Conclusion

The analysis highlights chest pain type and age as key predictors of heart disease. Decision trees and random forests provided valuable insights, with random forests offering better accuracy. This project demonstrates the integration of statistical and machine learning techniques to analyze and predict heart disease. Further research can explore additional predictors and external validation.