Heart Disease Analysis Case Study

FINDING WHICH FACTORS LEAD TO HEART DISEASE

About Heart Disease

Heart disease refers to the various conditions that hinder heart function. About 1 in 4 American deaths is linked to heart diseases. **The purpose of this study is to use machine learning to analyze and identify leading factors for heart disease.**

Heart Disease Risk Factors

This dataset contained data from 303 patients at a hospital. There were 14 variables, 13 of which were heart disease risk factors and the target variable (1 indicates heart disease and 0 indicates no heart disease).

After conducting background research on what causes heart disease, some of the common risk factors include:

- chest pains
- age
- sex (men are typically higher risk)
- family history
- smoking
- high blood pressure

Variables

1. **Age**
2. **Sex** (1 = Male, 0 = Female)
3. **Chest pain type** (4 values)
4. **Resting Blood Pressure**
5. **Serum cholesterol in mg/dl**
6. **Fasting blood sugar > 120 mg/dl**
7. **Resting electrocardiographic results** (3 values)
8. **The person's maximum heart rate achieved**
9. **Exercise induced angina**
10. **ST depression induced by exercise relative to rest**
11. **Slope of the peak exercise ST segment**.
12. **Number of major vessels** (0-3)
13. **Type of Thalassemia** - 3 = normal; 6 = fixed defect; 7 = reversible defect.
14. **Target** : 1 = Yes and 0 = No

Approach

Data Preprocessing

The variables were very detailed and contained more information than should be contained in one column. For this reason, data preprocessing included further separating the columns. Another preprocessing task was to introduce dummy variables to make all the data categorial.

Heart Disease Analysis Case Study

FINDING WHICH FACTORS LEAD TO HEART DISEASE

Approach

Building the Model
Utilized Random Forest Classifier to build the model. Graphed ROC curve and calculated AUC score to see model performance.

Finding Most Important Variables
Applied Permutation Importance to determine most important features.

Tools: Python (pandas, sklearn, matplotlib, seaborn, numpy, eli5)

92.4%

Accuracy in Predicting Heart Disease

Random Forest Model

Results

After combining the Permutation Importance chart with additional background research, the most important risk factors are thalassemia (reversible defect), number of major vessels, and maximum heart rate.

Most Important Risk Factors

Thalassemia (reversible defect)

Number of Major Vessels

Maximum Heart Rate