Project Proposal: Heart Disease Prediction

Introduction

Heart disease is one of the leading causes of death worldwide. Early detection and prediction of heart disease can significantly prolong patients' life. This project aims to develop a machine learning model to predict the presence of heart disease in individuals based on various health metrics.

Dataset and Data reading

```
[1]: import pandas as pd

     df = pd.read_csv('heart.csv',index_col=False)
```

This data set contains 14 attributes, including the predicted attribute. These 13 variables have been screened and are not affected by multicollinearity. The "target" field refers to the presence of heart disease in the patient.

Questions

- Can we predict the presence of heart disease based on these variables?
- We will use four models: logistic regression, K-NN, decision tree and SVM to study the data. Which machine learning model performs best in predicting heart disease in this dataset? Which hyperparameters perform best for that model?

Variables

- Target Variable:
  target: A binary variable where 1 indicates the presence of heart disease and 0 indicates its absence.
- Predictor Variables:

  - **age**
  - **sex**: 1 =male, 0 = female
  - **cp**: Chest pain type (4 types).
  - **trestbps**: Resting blood pressure (in mm Hg).
  - **chol**: Serum cholesterol in mg/dl.
  - **fbs**: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false).
  - **restecg**: Resting electrocardiographic results (values 0, 1, 2).
  - **thalach**: Maximum heart rate achieved.
  - **exang**: Exercise-induced angina (1 = yes; 0 = no).
  - **oldpeak**: ST depression induced by exercise relative to rest.
  - **slope**: The slope of the peak exercise ST segment.
  - **ca**: Number of major vessels (0-3) colored by fluoroscopy.
  - **thal**: A blood disorder indicator (3 = normal; 6 = fixed defect; 7 = reversible defect)

Methods

- Logistic Regression: Use this as a starting model and interpret the effect of each feature.
- Decision Tree: To explore a non-linear model and identify feature importance.
- K-Nearest Neighbors: To assess a simple distance-based approach.
- Support Vector Machine: To apply a robust classification algorithm for high-dimensional data.