

Report:

Assignment 1: Predicting Heart Disease

Your task is to predict the presence of heart disease in patients using data and AI. You will be provided with a dataset containing information about various attributes of patients such as age, sex, cholesterol levels, etc. along with the presence of heart disease (0 = no disease, 1-4 = varying degrees of disease). Your goal is to build a machine learning model that can accurately predict the presence of heart disease based on these attributes.

- Data set* <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Data Exploration:

In this section, performed various operations such as cleaning, feature selection, and feature engineering to prepare the data for analysis.

First, we will load the dataset and take a look at its features:

The dataset has 14 columns, including 13 features and the target variable. The features are:

1. age: Age of the patient in years
2. sex: Sex of the patient (1 = male; 0 = female)
3. cp: Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
4. trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
5. chol: Serum cholesterol level (in mg/dl)
6. fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. restecg: Resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
8. thalach: Maximum heart rate achieved
9. exang: Exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest
11. slope: The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12. ca: Number of major vessels (0-3) colored by fluoroscopy
13. thal: Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)

check for missing values in the dataset, then check the distribution of the target variable:

Data Analysis:

In this section, we will analyse the dataset and perform exploratory data analysis to identify any trends or patterns in the data.

First, we will plot the correlation matrix to see the correlation between the features:

Next, we split the dataset into training and testing sets using the `train_test_split()` function from scikit-learn. We then standardize the features using the `StandardScaler()` function to ensure that all features are on the same scale.

We train a logistic regression model using the `LogisticRegression()` class from scikit-learn and evaluate its performance using the accuracy score and classification report.