Project : Predicting Heart Disease

Problem Statement:

You are the data scientist at a medical research facility. The facility wants you to build a machine learning model to classify if the given data of a patient should tell if the patient is at the risk of a heart attack.

Heart Disease Dataset:

UCI Heart Disease Dataset
(https://archive.ics.uci.edu/ml/datasets/Heart+Disease?spm=5176.100239.blogcont54260.8.TRNGoO)

Lab Environment:

Jupyter Notebooks

Domain:

Healthcare

Tasks To Be Performed:

1. Data Analysis:
   a. Import the dataset
   b. Get information about the dataset (mean, max, min, quartiles etc.)
   c. Find the correlation between all fields

2. Data Visualization:
   a. Visualize the number of patients having a heart disease and not having a heart disease
   b. Visualize the age and whether a patient has disease or not
   c. Visualize correlation between all features using a heat map

3. Logistic Regression:
   a. Build a simple logistic regression model:
      i. Divide the dataset in 70:30 ratio
      ii. Build the model on train set and predict the values on test set
      iii. Build the confusion matrix and get the accuracy score

4. Decision Tree:
   a. Build a decision tree model:
      - i. Divide the dataset in 70:30 ratio
      - ii. Build the model on train set and predict the values on test set
      - iii. Build the confusion matrix and calculate the accuracy
      - iv. Visualize the decision tree using the Graphviz package

5. Random Forest:
   a. Build a Random Forest model:
      - i. Divide the dataset in 70:30 ratio
      - ii. Build the model on train set and predict the values on test set
      - iii. Build the confusion matrix and calculate the accuracy
      - iv. Visualize the model using the Graphviz package

6. Select the best model
   a. Print the confusion matrix of all classifiers
   b. Print the classification report of all classifiers
   c. Calculate Recall Precision and F1 score of all the models
   d. Visualize confusion matrix using heatmaps
   e. Select the best model based on the best accuracies