Task - 1: Implement a Machine Learning Model for the Classification problem.

Dataset link - <https://www.kaggle.com/datasets/ineubytes/heart-disease-dataset>

The Kaggle "Heart Disease Dataset" is a collection of information about people's hearts. It has details like age, gender, and medical measurements like blood pressure and cholesterol. This data helps doctors and scientists study heart disease and find ways to keep hearts healthy.

These are the columns of dataset:

Age: The age of the patient.

Sex: The gender of the patient (0 for female, 1 for male).

Chest Pain Type (cp): The type of chest pain experienced by the patient.

Resting Blood Pressure (trestbps): The resting blood pressure of the patient.

Serum Cholesterol Level (chol): The serum cholesterol level of the patient.

Fasting Blood Sugar (fbs): Fasting blood sugar level (> 120 mg/dl or <= 120 mg/dl).

Resting Electrocardiographic Results (restecg): Resting electrocardiographic results.

Maximum Heart Rate Achieved (thalach): The maximum heart rate achieved during exercise.

Exercise-Induced Angina (exang): Whether exercise-induced angina was present (0 for no, 1 for yes).

ST Depression (oldpeak): ST depression induced by exercise relative to rest.

Slope of the Peak Exercise ST Segment (slope): The slope of the peak exercise ST segment.

Number of Major Vessels Colored by Fluoroscopy (ca): The number of major vessels colored by fluoroscopy.

Thalassemia Type (thal): The type of thalassemia.

Presence of Heart Disease (target): The presence of heart disease (0 for no, 1 for yes).

Q/A for Task-1:

1. Which Python libraries did you find most useful in loading and exploring the dataset?

Ans: I found matplotlib, seaborn, numpy, pandas most useful in loading and exploring the dataset.

1. What preprocessing steps did you find necessary to apply to the heart dataset?

Ans: The preprocessing steps I found necessary to apply to the heart disease dataset are:

* 1. Removing Null Values
  2. Splitting Data into Train and Test Sets

1. What metrics were used to evaluate the Classification problem and why?

Ans: I used the confusion matrix, cross-validation, accuracy, overfitting, precision recall, F1 score metrics to evaluate the classification problem as they provide well-defined evaluation.

1. How did you detect overfitting in the model and what strategies did you use to mitigate it?

Ans: To detect overfitting, I used accuracy of train and test sets of data. If accuracy of train data is more than test data, then overfitting exists. If not, overfitting does not exists. To solve the overfitting problem I used 10-fold cross-validation method.