Problem Statement: Predicting Heart Disease Using Medical Data

Heart disease is one of the leading causes of mortality worldwide. Identifying individuals at risk of heart disease early can significantly improve health outcomes by facilitating timely intervention. The aim of this problem is to build a predictive model that can determine whether a person is likely to have heart disease based on various medical attributes.

Objective: To design and implement a machine learning model that predicts the likelihood of heart disease in individuals based on key health indicators, such as age, sex, cholesterol levels, resting blood pressure, and other relevant medical data.

Dataset: The dataset provided consists of patient records, each containing the following attributes:

1. **Age**: Age of the individual in years.
2. **Sex**: Gender of the individual (1 = male, 0 = female).
3. **Chest Pain Type**: Type of chest pain experienced (0-3, with different types of angina and no chest pain).
4. **Resting Blood Pressure**: Resting blood pressure in mm Hg.
5. **Cholesterol**: Serum cholesterol in mg/dl.
6. **Fasting Blood Sugar**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
7. **Resting Electrocardiographic Results**: Results of resting electrocardiography (values 0-2).
8. **Maximum Heart Rate Achieved**: The maximum heart rate achieved during a test.

9. **Exercise Induced Angina (exng)**: Presence of 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 (values 0-2).
12. **Number of Major Vessels (caa)**: The number of major vessels (0-4) colored by fluoroscopy.
13. **Thalassemia (thall)**: Blood disorder status (0-3).
14. **Output**: Target variable indicating the presence of heart disease (1 = disease, 0 = no disease).

Tasks:

1. Preprocess the dataset to handle missing values, scale numerical features, and encode categorical variables.
2. Split the dataset into training and testing sets.
3. Develop and train various machine learning models (e.g., logistic regression, decision trees, random forests, or neural networks) to predict the presence of heart disease.
4. Evaluate the models using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
5. Provide insights on the most significant medical attributes contributing to heart disease prediction.
6. Recommend potential areas for improving model performance or expanding the dataset for better generalization.

Key Challenges:

- Handling class imbalance if the number of individuals with and without heart disease is unequal.

- Dealing with multicollinearity between some medical attributes.
- Ensuring the model generalizes well to unseen patient data for real-world applicability.