Heart Disease Data

This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

Hint: heart_disease_uci.csv

Instructions:
----

1. Use Lifecycle of Data Sciece

Ans:

1.Load the dataset (heart_disease_uci.csv) and examine its structure and contents.

2.Analyze the distributions of the variables, check for missing values, and handle any data quality issues.

3.Explore the relationships between variables using visualizations and statistical summaries.

2. Use necessary data Preprocess techniques

Ans:

1. Split the dataset into features (independent variables) and target (dependent variable) columns.

2.Handle categorical variables, such as sex and chest pain type, by encoding them appropriately (e.g., one-hot encoding or label encoding).

3.Scale or normalize numerical features, such as age, resting blood pressure, and serum cholesterol, to ensure they are on similar scales.

4.Split the data into training and testing sets to evaluate model performance later.

3. Use various Regression and Classification techniques for comparision

Ans:

1.Choose a regression algorithm (e.g., linear regression, decision tree regression, random forest regression, etc.).

2.Fit the regression model using the training data and evaluate its performance on the testing data.

3.Use appropriate regression metrics (e.g., mean squared error, mean absolute error, R-squared) to assess the model's accuracy.

4.Compare the performance of different regression techniques to identify the best model.

4. Use metrics for regression and classification when needed.

Ans:

1.Select a classification algorithm (e.g., logistic regression, decision tree classification, random forest classification, etc.).

2.Train the classification model on the training data and evaluate its performance on the testing data.

3.Utilize classification metrics (e.g., accuracy, precision, recall, F1-score) to assess the model's performance.

4.Compare and contrast the results obtained from different classification techniques to determine the most effective model.

5. Use variosu Pipeline/Hyperparametr tuning techniques for improving performance

Ans:
1.Construct a data preprocessing pipeline that includes feature encoding, scaling, and any other necessary transformations.
2.Use techniques such as cross-validation and grid search to optimize the hyperparameters of the selected regression and classification models.
3.Compare the performance of the tuned models with the initial models and identify the best-performing models for each task.