**HEART DISEASE PREDICTION USING ML**

The problem statement revolves around creating a predictive model to determine the likelihood of heart disease in individuals based on a set of clinical and demographic attributes. Our data set consists of necessary components like Age, blood pressure, maximum heart rate and other 11 components to analyze whether the person suffers with heart disease or not. We indicated binary variable 1 for yes and 0 for no. Our data set contains 270 rows and 14 columns.

In this prediction we selected four machine learning models to find the effective model for our prediction. We selected logistic Regression, Random Forest, Decision Tree and K-Nearest neighbor as our testing and training models. We only selected these models :

**K-Nearest Neighbor:** We chose this because the data distribution is not highly complex and it makes better predictions and classifications.

**Logistic Regression:** We chose this because it clearly tells the relationship between the independent variables and likely hood of a particular outcome.

**Random Forest:** It combines multiple decision trees to improve prediction performance which is important for our data set as we require high accuracy.

**Decision Tree:** It can handle both classification and regression tasks which is useful for capturing non-linear relationships in the data.

Exploratory Data Analysis:

Exploratory Data Analysis or EDA is used to take insights from the data. The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. It allows to understand the data better before making any assumption. It provides the context needed to develop an appropriate model. This involves checking for missing values, summarizing the statistics, and creating data visualizations to gain insights into the dataset.

C-map : This is a cmap which calculates the summary statistics of our data frame.It calculates mean, median, standard deviation, highest value and lowest value which helps in understanding the characteristics of the datset In this the darker color represents highest value.

Box Plot : This represents box plot of our dataset which is used to identify any potential outliers. The main use of a boxplot in code is to provide a visual summary of the data distribution.

Histogram : Histograms are particularly useful when analyzing large datasets, as they allow you to quickly identify patterns and trends in the data. They are also useful for comparing the distribution of data between different groups or datasets.

Pair plot graph : This is used to visualize the pairwise relationships between multiple variables in a dataset. Pair plot graphs are a powerful tool for exploratory data analysis, providing a visual representation of the relationships between variables and aiding in the identification of patterns, trends, correlations, and outliers in the data.

Splitting Independent and dependent Variables :

The independent and dependent variables are separated, where X contains the independent features, and y contains the target variable.

Correlation Matrix:

This section calculates the correlation matrix between features and visualizes it using a heatmap. The resulting heatmap visually represents the correlations between different features in the dataset. Positive correlations are typically indicated by lighter colors, while negative correlations are shown as darker colors. This visualization can help in feature selection, understanding relationships in the data, and making informed decisions during data analysis and modeling.

Data Splitting:

Here we are dividing the dataset into training data and testing data. 80% of the dataset is used for training and 20% of data is used for testing.

Data Preprocessing:

Standard scaling is applied to certain features using Standard Scaler to normalize the data. Standardize numerical features to have a mean of 0 and standard deviation of 1.

Handling Outliers: Handling outliers is a process of identifying and dealing with extreme values in a dataset that are significantly different from the majority of the data points. Identify and limit extreme values (outliers) in each feature. Handling outliers is important because they can significantly impact statistical analysis and modeling results. Outliers can skew the distribution, affect measures of central tendency and dispersion, and distort relationships between variables.

Modeling and Evaluation:

Four different machine learning models are imported: K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, and Random Forest. These models are trained using the training data with fit. The accuracy of each model is evaluated on the test data using accuracy\_score and printed.

The plot\_classification\_report function is a custom function that generates classification reports for the training and test sets of each model. Classification reports provide detailed information about the performance of classification models, including metrics such as precision, recall, F1-score, and support for each class. By calling this function for each model and providing the appropriate arguments, the code generates and displays these reports to assess the performance of each model on the training and test data. This helps in evaluating and comparing the models in terms of their predictive accuracy and other classification metrics.

Confusion matrices are an essential tool for evaluating the performance of classification models, as they provide insights into true positives, true negatives, false positives, and false negatives. These matrices help in understanding the quality of predictions made by each model and can be particularly useful for understanding classification errors.

In conclusion, this code offers a comprehensive solution for heart disease prediction through machine learning. While Logistic Regression proves to be the most effective model