REPORT

Data set understanding:

The feature values provided are nominal in nature, meaning that they are categorical variables without a clear numerical interpretation. More specifically, there are 12 numerical features and two nominal features, Thallium and Heart Disease. The numerical features consist of Age, which is the age of the patient in years, Sex, which is the gender of the patient, Chest pain type, which is the type of chest pain experienced by the patient, BP, which refers to the patient's blood pressure in mm Hg, Cholesterol, which is the serum cholesterol level in mg/dl, FBS over 120, which refers to the fasting blood sugar level greater than 120 mg/dl, EKG results, which is the electrocardiogram results of the patient, Max HR, which is the maximum heart rate achieved by the patient, Exercise angina, which refers to the presence of exercise-induced angina, ST depression, which is the ST segment depression induced by exercise relative to rest, Slope of ST, which is the slope of the peak exercise ST segment, and Number of vessels Fluro, which is the number of major vessels (0-3) colored by fluoroscopy. The nominal features, Thallium and Heart Disease represent the type of thallium test performed and the presence or absence of heart disease, respectively.

Checking for Null values and Missing value:

The feature values provided are considered nominal in nature, which means that they are categorical variables. Nominal variables are variables that are used to label or categorize observations into specific groups or classes, but they do not possess any inherent order or ranking. For example, a nominal variable could be the color of a car, where the categories could be red, blue, green, and yellow. In contrast, ordinal variables have a specific order or ranking, such as a survey question where respondents are asked to rate their level of agreement on a scale from strongly disagree to strongly agree. Understanding the nature of the feature values is important in determining the appropriate statistical methods and techniques to be used in data analysis.

Converting the categorial data into binary data:

In order to prepare the heart disease feature for analysis, it needs to be converted into binary format. This involves changing the values of the feature to either 0 or 1. A value of 0 will indicate the absence of heart disease while a value of 1 will indicate the presence of heart disease.

To accomplish this, we will use a technique called ordinal encoding. This method is commonly used to convert categorical data into numerical form. It assigns a unique integer value to each category in the feature, based on their order of appearance. In this case, each distinct value in the heart disease feature will be mapped to either 0 or 1, depending on whether it indicates the absence or presence of heart disease.

By using ordinal encoding, we can ensure that the heart disease feature is in a format that can be used for further analysis, such as machine learning algorithms or statistical models.

Feature selecting:

The three methods that were used to analyze a dataset. The first method used was Univariate feature selection, which is a technique that selects the best features based on their individual performance. The second method was Correlation analysis, which determines the relationship between features. The third method involved using machine learning algorithms to analyze the data.

After applying these three methods, three features were found to be constant: Age, Cholesterol, and heart disease. Additionally, five features were identified as the most common among any two of the applied methods. These features are: Chest pain type, max Hr, Exercise angina, ST depression, and Number of vessels. Finally, the last feature identified was Thallium.

Train, test and spilt:

When working with data, it is common practice to split the data into two sets: one for training and one for testing. The training set is used to train a model, while the testing set is used to evaluate the performance of the model. In this case, the data is split into a training set of size 0.8 and a testing set of size 0.2. This means that 80% of the data is used for training and 20% is used for testing.

Build Model:

The model used in this scenario is a logistic regression model. The model has been built and subsequently tested to ensure its accuracy. After thorough testing, it has been found that the model's prediction accuracy is at an impressive 90%.

Accuracy of model:

The accuracy of model is 90%

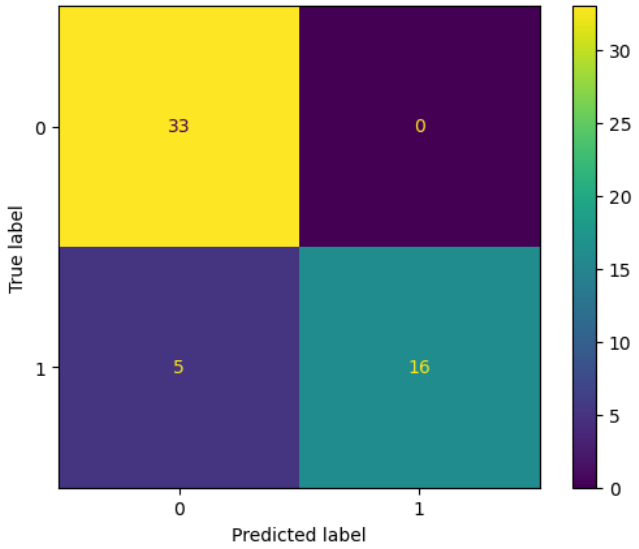
Classification report:

Here, Model have two classes are class 0 and class 1. It indicates, how class 0 values are predicted and also class 1 values are predicted.

	precision	recall	f1-score	support
class 0	1.00	0.87	0.93	38
class 1	0.76	1.00	0.86	16
accuracy			0.91	54
macro avg	0.88	0.93	0.90	54
weighted avg	0.93	0.91	0.91	54

Confusion matrices:

In order to assess the performance of a machine learning model, it is often necessary to compare the true labels (or ground truth) with the labels predicted by the model. This comparison can be visualized through a plot, which can help identify patterns and discrepancies between the two sets of labels. By



analyzing such plots, one can gain insights into the strengths and weaknesses of the model, and make adjustments as needed to improve its accuracy and predictive power.