MACHINE LEARNING

Course-End Project: Healthcare

PROJECT EXPLANATION

TASK-1 (PRELIMINARY ANALYSIS)

- Firstly import the health-care dataset from the root directory using pd.excel() function and store it in a dataframe object

- By using the head() and tail() function on the Dataframe we have checked the top 5 and bottom 5 rows of the dataset respectively.

- Using the shape operation we got to know the no of rows and columns present in the dataframe.

- To find the missing values and complete information about the dataset we have used info() function and isnull().sum() onto the Dataframe and found that we do not have any missing values in the dataset and therefore missing values treatment is not needed now.

TASK-2 (EXPLORATORY DATA ANALYSIS OF HEALTHCARE DATA)

- We have used the countplot to count the number of true values for CVD and non-CVD'S for the given data and infer that true CVD values are more than non-CVD values

- To draw the conclusion between the number of Target values(Dependent variable) vs the sex variable we made use of bar graph and from the bar graph we can say that the males are 93 and females are 72 and by this conclude males are more prone to CVD

- To draw the conclusion between the number of Target values(Dependent variable) vs the chest pain (CPL) variable we made use of bar graph and from the bar graph we can Infer that 69 people are suffering from non-anginal type of chest pain are detected with CVD, Whereas the 104 people who are suffering from the Asymptomatic type of chest pain are less likely to detect the CVD.

- To draw the conclusion between the number of Target values(Dependent variable) vs the RESTECG variable we made use of the bar graph and from the bar graph we can say that the Abnormal resting electrocardiographic results show maximum number of occurrences of CVD.

- By using the histogram to analyze the variables blood pressure(trestbps) and target variable , we can infer that people whose blood pressure varies between 120 to 140 are more likely to detect the CVD.

- Last part of the EDA we use the pair plot which shows the groups of scatter plots telling the relationship with each variable with the remaining other variables in the dataset .

TASK-3(MODELING OF THE HEALTHCARE DATA)

- Firstly we use logistic regression because ultimately we are dealing with the classification problem where we are going to predict the likelihood of heart attack using the patient data .

- Dataset is splitted into the train data and the test data with 75;25 ratio where 75 percent is the training data and 25 percent is the testing data .

- We have imported the Logistic regression model from a model called linear_model .

- Now the train data and test data is being fit into the model object .

- After using the test data to realize the ability of the model towards the trained data we have acquired the accuracy of prediction is nearly about 84.21 percent

- The above same process has been repeated to the dataset but now we are using a Random forest model .

- ‘The accuracy achieved now in this case for predicting heart attacks is nearly about 85.52 percent.

- The main inference here is that the dataset is split into the standard 75;25 ratio and prediction is done but incase if we use the ratio of splitting to be 80:20
Then the accuracy of the prediction of heart attacks might still go up compared to 85.52 percent.