Day 38

DIY

Q1. Problem Statement: Logistic Regression

You are given a categorical dataset – *"Heart_Disease.csv."* Load the dataset into a DataFrame. Considering the *"TenYearCHD"* column as the target variable, perform the following tasks:

1. Explore the *"Heart_Disease.csv"* dataset, identify the null values and fill them with the mean value of their respective columns
2. Split the data into test and train parts using `train_test_split()` function in 80:20 ratio (80% train, 20% test)
3. Perform scaling of numeric data using the `StandardScaler()` function
4. Build a Logistic regression model using the test dataset and test the model using the test dataset
5. Print the classification report of the model
6. Calculate the confusion matrix and plot the same using a heatmap
7. Calculate and print the accuracy score of the model
8. Print the decision boundary for θ = 0, θ = 1 and 2

Dataset:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Sample Output:

1. Explore the *"Heart_Disease.csv"* dataset, identify the null values and fill them with the mean value of their respective columns

```
male                0
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
TenYearCHD          0
dtype: int64
```

```
male              0
age               0
education         0
currentSmoker     0
cigsPerDay        0
BPMeds            0
prevalentStroke   0
prevalentHyp      0
diabetes          0
totChol           0
sysBP             0
diaBP             0
BMI               0
heartRate         0
glucose           0
TenYearCHD        0
dtype: int64
```

2. Split the data into test and train parts using train_test_split() function in 80:20 ratio (80% train, 20% test)

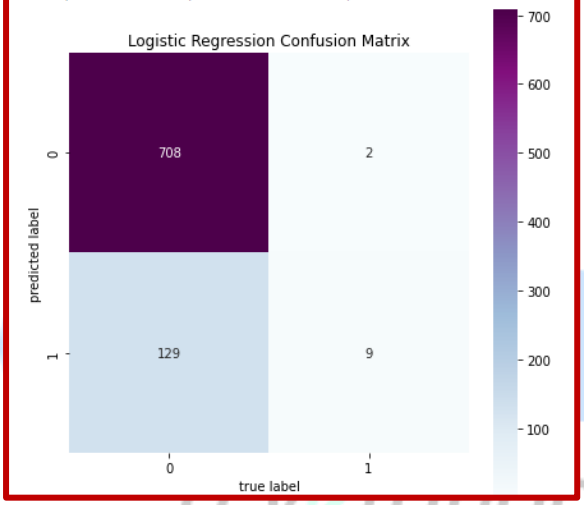
```
After splitting the data-
size of input train data is: 433952
sizeof input test data is: 108576
size of output train data is: 54272
size of output test data is: 13600
```

3. Perform scaling of numeric data using the StandardScaler() function
4. Build a Logistic regression model using the test dataset and test the model using the test dataset
5. Print the classification report of the model

	precision	recall	f1-score	support
0	0.85	1.00	0.92	710
1	0.82	0.07	0.12	138
accuracy			0.85	848
macro avg	0.83	0.53	0.52	848
weighted avg	0.84	0.85	0.79	848

6. Calculate the confusion matrix and plot the same using a heatmap

```
Confusion matrix is
 [[708   2]
 [129   9]]
Text(42.0, 0.5, 'predicted label')
```



7. Calculate and print the accuracy score of the model

```
accuracy score :  0.8455188679245284
accuracy: 85 %
```

8. Print the decision boundary for θ = 0, θ = 1 and

```
[-1.99450414]
[[ 0.21635451  0.52284324 -0.00373038  0.01763908  0.27002551  0.01692892
   0.08452846  0.16742972  0.054612    0.0912552   0.29073846 -0.09032936
   0.03575889 -0.01982821  0.12616453]]
```