In [1]:
```
#Aim : To perform and find the accuracy of Logistic Regression
```

In [2]:
```
#Name: Shruti Anil Dhote
#RolL no. :72
#Sub : ET1
#section:C
#Date:16/08/2024
```

In [3]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')
```

In [4]:
```
import os
```

In [5]:
```
os.getcwd()
```

Out[5]:
```
'C:\\Users\\SURUTI DHOTE'
```

In [6]:
```
os.chdir("C:\\Users\\SURUTI DHOTE\\Desktop")
```

In [7]:
```
df=pd.read_csv("framingham.csv")
```

In [8]:
```
#The "Framingham" heart disease dataset includes over 4,240 records, 15 attributes.
#The goal of the dataset is to predict whether the patient has 10-year risk of futu
```

In [9]:
```
df.head()
```

Out[9]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	dia
0	1	39	4.0	0	0.0	0.0	0	0	
1	0	46	2.0	0	0.0	0.0	0	0	
2	1	48	1.0	1	20.0	0.0	0	0	
3	0	61	3.0	1	30.0	0.0	0	1	
4	0	46	3.0	1	23.0	0.0	0	0	

In [10]:
```
df.describe()
```

Out[10]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalent
count	4240.000000	4240.000000	4135.000000	4240.000000	4211.000000	4187.000000	4240.0
mean	0.429245	49.580189	1.979444	0.494104	9.005937	0.029615	0.0
std	0.495027	8.572942	1.019791	0.500024	11.922462	0.169544	0.0
min	0.000000	32.000000	1.000000	0.000000	0.000000	0.000000	0.0
25%	0.000000	42.000000	1.000000	0.000000	0.000000	0.000000	0.0
50%	0.000000	49.000000	2.000000	0.000000	0.000000	0.000000	0.0
75%	1.000000	56.000000	3.000000	1.000000	20.000000	0.000000	0.0
max	1.000000	70.000000	4.000000	1.000000	70.000000	1.000000	1.0

In [11]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4240 entries, 0 to 4239
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   male             4240 non-null   int64  
 1   age              4240 non-null   int64  
 2   education        4135 non-null   float64
 3   currentSmoker    4240 non-null   int64  
 4   cigsPerDay       4211 non-null   float64
 5   BPMeds           4187 non-null   float64
 6   prevalentStroke  4240 non-null   int64  
 7   prevalentHyp     4240 non-null   int64  
 8   diabetes         4240 non-null   int64  
 9   totChol          4190 non-null   float64
 10  sysBP            4240 non-null   float64
 11  diaBP            4240 non-null   float64
 12  BMI              4221 non-null   float64
 13  heartRate        4239 non-null   float64
 14  glucose          3852 non-null   float64
 15  TenYearCHD       4240 non-null   int64  
dtypes: float64(9), int64(7)
memory usage: 530.1 KB
```

In [12]:
```
df.isna().sum()
```

Out[12]:

```
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
```

In [13]:
```
#Since, only a few rows have null values in them, we are only removing those rows f
#df = df.dropna(subset=['heartRate','BMI','cigsPerDay','totChol','BPMeds'])
```

In [14]:
```
df
```

Out[14]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp
0	1	39	4.0	0	0.0	0.0	0	0
1	0	46	2.0	0	0.0	0.0	0	0
2	1	48	1.0	1	20.0	0.0	0	0
3	0	61	3.0	1	30.0	0.0	0	1
4	0	46	3.0	1	23.0	0.0	0	0
...	...	...	...	...	...	...	...	...
4235	0	48	2.0	1	20.0	NaN	0	0
4236	0	44	1.0	1	15.0	0.0	0	0
4237	0	52	2.0	0	0.0	0.0	0	0
4238	1	40	3.0	0	0.0	0.0	0	1
4239	0	39	3.0	1	30.0	0.0	0	0

4240 rows × 16 columns

MISSING VALUE TREATMENT

Since, 'glucose' and 'education' columns had a significant amount of null values, so we replaced them with the mean of values for theirrespective columns

In [15]:
```
df['glucose'].fillna(value = df['glucose'].mean(),inplace=True)
```

In [16]:
```
df['education'].fillna(value = df['education'].mean(),inplace=True)
```

In [17]:
```
df['heartRate'].fillna(value = df['heartRate'].mean(),inplace=True)
```

In [18]:
```
df['BMI'].fillna(value = df['BMI'].mean(),inplace=True)
```

In [19]:
```
df['cigsPerDay'].fillna(value = df['cigsPerDay'].mean(),inplace=True)
```

In [20]:
```
df['totChol'].fillna(value = df['totChol'].mean(),inplace=True)
```

In [21]:
```
df['BPMeds'].fillna(value = df['BPMeds'].mean(),inplace=True)
```

In [22]:
```
df.isna().sum()
```

Out[22]:
```
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
```

In [23]:
```
#Splitting the dependent and independent variables.
x = df.drop("TenYearCHD",axis=1)
y = df['TenYearCHD']
```

In [24]:
```
x #checking the features
```

Out[24]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp
0	1	39	4.0	0	0.0	0.000000	0	0
1	0	46	2.0	0	0.0	0.000000	0	0
2	1	48	1.0	1	20.0	0.000000	0	0
3	0	61	3.0	1	30.0	0.000000	0	1
4	0	46	3.0	1	23.0	0.000000	0	0
...	...	...	...	...	...	...	...	...
4235	0	48	2.0	1	20.0	0.029615	0	0
4236	0	44	1.0	1	15.0	0.000000	0	0
4237	0	52	2.0	0	0.0	0.000000	0	0
4238	1	40	3.0	0	0.0	0.000000	0	1
4239	0	39	3.0	1	30.0	0.000000	0	0

4240 rows × 15 columns

TRAIN TEST SPLIT

In [25]:
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=42)
```

In [26]:
```
y_train
```

Out[26]:
```
1427    0
3257    0
3822    0
1263    0
3575    0
       ..
3444    0
466     0
3092    0
3772    0
860     0
Name: TenYearCHD, Length: 3392, dtype: int64
```

Logistic Regression Algorithm

In [27]:
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression().fit(x_train,y_train)
model.score(x_train, y_train)
```

Out[27]:
```
0.8484669811320755
```

In [ ]:
```

```