In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [2]:

```
data=pd.read_csv(r"C:\Users\joel\Downloads\framingham.csv")
data
```

Out[2]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalent
0	1	39	4.0	0	0.0	0.0	0	
1	0	46	2.0	0	0.0	0.0	0	
2	1	48	1.0	1	20.0	0.0	0	
3	0	61	3.0	1	30.0	0.0	0	
4	0	46	3.0	1	23.0	0.0	0	
...	...	...	...	...	...	...	...	
4233	1	50	1.0	1	1.0	0.0	0	
4234	1	51	3.0	1	43.0	0.0	0	
4235	0	48	2.0	1	20.0	NaN	0	
4236	0	44	1.0	1	15.0	0.0	0	
4237	0	52	2.0	0	0.0	0.0	0	

4238 rows × 16 columns

In [3]:

```
data.head()
```

Out[3]:

	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

In [4]:

```
data.tail()
```

Out[4]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalent
4233	1	50	1.0	1	1.0	0.0	0	
4234	1	51	3.0	1	43.0	0.0	0	
4235	0	48	2.0	1	20.0	NaN	0	
4236	0	44	1.0	1	15.0	0.0	0	
4237	0	52	2.0	0	0.0	0.0	0	

In [5]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   male             4238 non-null   int64  
 1   age              4238 non-null   int64  
 2   education        4133 non-null   float64
 3   currentSmoker    4238 non-null   int64  
 4   cigsPerDay       4209 non-null   float64
 5   BPMeds           4185 non-null   float64
 6   prevalentStroke  4238 non-null   int64  
 7   prevalentHyp     4238 non-null   int64  
 8   diabetes         4238 non-null   int64  
 9   totChol          4188 non-null   float64
 10  sysBP            4238 non-null   float64
 11  diaBP            4238 non-null   float64
 12  BMI              4219 non-null   float64
 13  heartRate        4237 non-null   float64
 14  glucose          3850 non-null   float64
 15  TenYearCHD       4238 non-null   int64  
dtypes: float64(9), int64(7)
memory usage: 529.9 KB
```

In [6]:

```
data.shape
```

Out[6]:

```
(4238, 16)
```

In [7]:

```
data.describe()
```

Out[7]:

education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diat
4133.000000	4238.000000	4209.000000	4185.000000	4238.000000	4238.000000	4238.00
1.978950	0.494101	9.003089	0.029630	0.005899	0.310524	0.02
1.019791	0.500024	11.920094	0.169584	0.076587	0.462763	0.15
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
3.000000	1.000000	20.000000	0.000000	0.000000	1.000000	0.00
4.000000	1.000000	70.000000	1.000000	1.000000	1.000000	1.00

In [8]:

```
data.isnull().any()
```

Out[8]:

```
male               False
age                False
education           True
currentSmoker      False
cigsPerDay          True
BPMeds              True
prevalentStroke    False
prevalentHyp       False
diabetes           False
totChol             True
sysBP              False
diaBP              False
BMI                 True
heartRate           True
glucose             True
TenYearCHD         False
dtype: bool
```

In [35]:

```
data.isnull().sum()
```

Out[35]:

```
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 [36]:

```


data['TenYearCHD'].value_counts()
```

Out[36]:

```
TenYearCHD
0    3594
1     644
Name: count, dtype: int64
```

In [24]:

```
x=data.drop(columns='TenYearCHD',axis=1)
y=data['TenYearCHD']
```

In [25]:

```
print(x)
```

```
      male  age  education  currentSmoker  cigsPerDay  BPMeds  \
0        1   39        4.0              0         0.0     0.0   
1        0   46        2.0              0         0.0     0.0   
2        1   48        1.0              1        20.0     0.0   
3        0   61        3.0              1        30.0     0.0   
4        0   46        3.0              1        23.0     0.0   
...    ...  ...        ...            ...         ...     ...   
4233     1   50        1.0              1         1.0     0.0   
4234     1   51        3.0              1        43.0     0.0   
4235     0   48        2.0              1        20.0     NaN   
4236     0   44        1.0              1        15.0     0.0   
4237     0   52        2.0              0         0.0     0.0   

      prevalentStroke  prevalentHyp  diabetes  totChol  sysBP  diaBP    BM
I
0                   0             0         0    195.0  106.0   70.0  26.9
7  \
1                   0             0         0    250.0  121.0   81.0  28.7
3
2                   0             0         0    245.0  127.5   80.0  25.3
4
3                   0             1         0    225.0  150.0   95.0  28.5
8
4                   0             0         0    285.0  130.0   84.0  23.1
0
...               ...           ...       ...      ...    ...    ...
...
4233                0             1         0    313.0  179.0   92.0  25.9
7
4234                0             0         0    207.0  126.5   80.0  19.7
1
4235                0             0         0    248.0  131.0   72.0  22.0
0
4236                0             0         0    210.0  126.5   87.0  19.1
6
4237                0             0         0    269.0  133.5   83.0  21.4
7

      heartRate  glucose
0          80.0     77.0
1          95.0     76.0
2          75.0     70.0
3          65.0    103.0
4          85.0     85.0
...         ...      ...
4233       66.0     86.0
4234       65.0     68.0
4235       84.0     86.0
4236       86.0      NaN
4237       80.0    107.0

[4238 rows x 15 columns]
```

In [26]:

```
print(y)
```

```
0       0
1       0
2       0
3       1
4       0
       ..
4233    1
4234    0
4235    0
4236    0
4237    0
Name: TenYearCHD, Length: 4238, dtype: int64
```

In [27]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=2,stratify=y,random_state=2
```

In [28]:

```
print(x.shape,x_train.shape,x_test.shape)
```

```
(4238, 15) (4236, 15) (2, 15)
```