In [17]:
```
#import libraries
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 [18]:
```
df=pd.read_csv(r"C:\Users\Mastan Reddy\Downloads\archive (6).zip")
df
```

Out[18]:

e	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	t
l	39	4.0	0	0.0	0.0	0	0	0	
)	46	2.0	0	0.0	0.0	0	0	0	
l	48	1.0	1	20.0	0.0	0	0	0	
)	61	3.0	1	30.0	0.0	0	1	0	
)	46	3.0	1	23.0	0.0	0	0	0	
.	...	...	...	...	...	...	...	...	
l	50	1.0	1	1.0	0.0	0	1	0	
l	51	3.0	1	43.0	0.0	0	0	0	
)	48	2.0	1	20.0	NaN	0	0	0	
)	44	1.0	1	15.0	0.0	0	0	0	
)	52	2.0	0	0.0	0.0	0	0	0	

× 16 columns

In [19]:
```
df.head()
```

Out[19]:

	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 [20]:
```
df.tail()
```

Out[20]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp
4233	1	50	1.0	1	1.0	0.0	0	1
4234	1	51	3.0	1	43.0	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

In [21]:
```
df.describe
```

Out[21]:
```
<bound method NDFrame.describe of       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    BMI
\
0                   0             0         0    195.0  106.0   70.0  26.97
1                   0             0         0    250.0  121.0   81.0  28.73
2                   0             0         0    245.0  127.5   80.0  25.34
3                   0             1         0    225.0  150.0   95.0  28.58
4                   0             0         0    285.0  130.0   84.0  23.10
...               ...           ...       ...      ...    ...    ...    ...
4233                0             1         0    313.0  179.0   92.0  25.97
4234                0             0         0    207.0  126.5   80.0  19.71
4235                0             0         0    248.0  131.0   72.0  22.00
4236                0             0         0    210.0  126.5   87.0  19.16
4237                0             0         0    269.0  133.5   83.0  21.47

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

[4238 rows x 16 columns]>
```

In [22]:
```
df.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 [23]:
```
df.shape
```

Out[23]:
```
(4238, 16)
```

In [24]:
```
df.isnull().sum()
```

Out[24]:
```
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 [30]:
```
df['education'].value_counts()
```

Out[30]:
```
1.0    1720
2.0    1253
3.0     687
4.0     473
Name: education, dtype: int64
```

In [29]:
```
df['age'].value_counts()
```

Out[29]:
```
40    191
46    182
42    180
41    174
48    173
39    169
44    166
45    162
43    159
52    149
51    146
55    145
38    144
47    141
50    140
53    139
49    132
54    132
56    123
57    123
59    119
58    117
60    111
61    110
63    110
62     99
64     93
37     92
36     84
65     57
67     45
35     42
66     38
34     18
68     18
69      7
33      5
70      2
32      1
Name: age, dtype: int64
```

In [ ]:

In [ ]:

In [ ]: