KNN CLASSIFIER

In [2]:
```
#Name : Shreya Sharma
#Roll no. : 46
#Sectin : 3B
#Date : 08/10/2024
```

In [2]:
```
#Aim : To perform operation on knn classifier
```

In [1]:
```
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 [2]:
```
import os
```

In [3]:
```
os.getcwd()
```

Out[3]:
```
'C:\\Users\\pravi'
```

In [4]:
```
os.chdir("C:\\Users\\pravi\\Desktop")
```

In [5]:
```
df=pd.read_csv("framingham.csv")
```

In [6]:
```
df.head()
```

Out[6]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23

In [12]:
```
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 [15]:
```
df.isna().sum()
```

Out[15]:
```
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 [17]:
```
df
```

Out[17]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0
...	...	...	...	...	...	...	...	...	...	...	...	...
4233	1	50	1.0	1	1.0	0.0	0	1	0	313.0	179.0	92.0
4234	1	51	3.0	1	43.0	0.0	0	0	0	207.0	126.5	80.0
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0

4238 rows × 16 columns

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

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

In [23]:
```
df['heartRate'].fillna(value = df['heartRate'].mean(),inplace=True)
```

In [25]:
```
df['BMI'].fillna(value = df['BMI'].mean(),inplace=True)
```

In [27]:
```
df['cigsPerDay'].fillna(value = df['cigsPerDay'].mean(),inplace=True)
```

In [29]:
```
df['totChol'].fillna(value = df['totChol'].mean(),inplace=True)
```

In [31]:
```
df['BPMeds'].fillna(value = df['BPMeds'].mean(),inplace=True)
```

In [33]:
```
df.isna().sum()
```

Out[33]:
```
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 [35]:
```
x = df.drop("TenYearCHD",axis=1)
y = df['TenYearCHD']
```

In [37]:
```
x
```

Out[37]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
0	1	39	4.0	0	0.0	0.00000	0	0	0	195.0	106.0	70.0
1	0	46	2.0	0	0.0	0.00000	0	0	0	250.0	121.0	81.0
2	1	48	1.0	1	20.0	0.00000	0	0	0	245.0	127.5	80.0
3	0	61	3.0	1	30.0	0.00000	0	1	0	225.0	150.0	95.0
4	0	46	3.0	1	23.0	0.00000	0	0	0	285.0	130.0	84.0
...	...	...	...	...	...	...	...	...	...	...	...	...
4233	1	50	1.0	1	1.0	0.00000	0	1	0	313.0	179.0	92.0
4234	1	51	3.0	1	43.0	0.00000	0	0	0	207.0	126.5	80.0
4235	0	48	2.0	1	20.0	0.02963	0	0	0	248.0	131.0	72.0
4236	0	44	1.0	1	15.0	0.00000	0	0	0	210.0	126.5	87.0
4237	0	52	2.0	0	0.0	0.00000	0	0	0	269.0	133.5	83.0

4238 rows × 15 columns

In [39]:
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=42)
```

In [41]:
```
x_train
```

Out[41]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
3252	1	40	4.0	1	30.0	0.0	0	0	0	205.0	131.0	81.0
3946	0	57	2.0	0	0.0	0.0	0	1	0	250.0	152.5	92.5
1261	0	47	1.0	0	0.0	0.0	0	0	0	230.0	123.0	71.0
2536	1	41	2.0	1	30.0	0.0	0	0	0	228.0	113.0	82.5
4089	0	64	1.0	0	0.0	0.0	0	1	0	232.0	149.5	84.0
...	...	...	...	...	...	...	...	...	...	...	...	...
3444	0	36	1.0	1	5.0	0.0	0	1	0	222.0	147.0	94.0
466	0	57	3.0	1	15.0	0.0	0	0	0	250.0	125.0	74.0
3092	0	60	2.0	0	0.0	0.0	0	1	0	298.0	133.0	89.0
3772	1	39	2.0	1	10.0	0.0	0	0	0	215.0	102.0	64.5
860	0	35	2.0	0	0.0	0.0	0	0	0	248.0	107.0	73.0

3390 rows × 15 columns

In [43]:
```
y_train
```

Out[43]:
```
3252    0
3946    0
1261    0
2536    0
4089    0
       ..
3444    0
466     0
3092    0
3772    0
860     0
Name: TenYearCHD, Length: 3390, dtype: int64
```

In [45]:
```
x_test
```

Out[45]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
3188	1	63	1.0	0	0.0	0.0	0	1	0	190.0	148.0	90.0
764	1	45	3.0	0	0.0	0.0	0	0	0	162.0	125.0	89.0
3264	0	51	1.0	1	2.0	0.0	0	0	0	261.0	127.0	81.0
1967	1	45	3.0	1	30.0	0.0	0	0	0	250.0	126.0	89.5
2185	0	45	2.0	1	3.0	0.0	0	0	0	250.0	130.0	80.0
...	...	...	...	...	...	...	...	...	...	...	...	...
3303	1	47	1.0	0	0.0	0.0	0	0	0	259.0	139.0	79.0
4056	1	44	2.0	0	0.0	0.0	0	0	0	254.0	130.0	80.0
4210	1	50	1.0	0	0.0	0.0	0	0	0	282.0	126.5	88.0
3971	1	64	3.0	0	0.0	0.0	0	1	1	195.0	176.0	78.0
2540	1	55	3.0	1	20.0	0.0	0	0	0	214.0	110.0	71.0

848 rows × 15 columns

In [47]:
```
y_test
```

Out[47]:
```
3188    0
764     0
3264    0
1967    0
2185    0
       ..
3303    1
4056    0
4210    0
3971    0
2540    0
Name: TenYearCHD, Length: 848, dtype: int64
```

In [49]:
```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5, p=2, metric='minkowski')
knn.fit(x_train, y_train)
acc = knn.score(x_test,y_test)*100
print(acc)
```

```
83.13679245283019
```