cvd-prediction-1

April 21, 2024

CONTENTS

INTRODUCTION

• The goal of this project is to create machine learning models to predict whether a person has heart disease based on various health and lifestyle factors.

• The dataset includes various features related to patients' such as health,lifestyle including age, sex, checkup frequency, exercise habits, smoking history, and the presence of various diseases.

DATASET : Cardiovascular Diseases Risk Prediction Dataset

PROCEDURE

1. Data Loading and Preprocessing: Load the data and preprocess it for analysis and modeling.
2. Exploratory Data Analysis (EDA): Perform exploratory data analysis to gain insights into the dataset understand the distributions of features, and explore potential relationships between the features and the outcome.
3. Data cleaning : We clean and preprocess the data to prepare it for machine learning. This included handling missing values, encoding categorical variables, and scaling numerical variables.
4. Model Training and Validation: Train the model using a train-test split strategy and make predictions on the test set
5. Model Evaluation: Evaluate the performance of the trained model using appropriate algorithms.

IMPORTING REQUIRED LIBRARIES

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

IMPORTING DATASET

```
[ ]: df=pd.read_csv('/content/heartdata.csv')
     df
```

```
[ ]:        General_Health                 Checkup Exercise Skin_Cancer  \
    0                Poor  Within the past 2 years       No          No
    1           Very Good     Within the past year       No          No
    2           Very Good     Within the past year      Yes          No
    3                Poor     Within the past year      Yes          No
    4                Good     Within the past year       No          No
    …                 …                        …      …         …
    308847      Very Good     Within the past year      Yes          No
    308848           Fair  Within the past 5 years      Yes          No
    308849      Very Good      5 or more years ago      Yes          No
    308850      Very Good     Within the past year      Yes          No
    308851      Excellent     Within the past year      Yes          No

           Other_Cancer Depression                                    Diabetes  \
    0                No         No                                         No
    1                No         No                                        Yes
    2                No         No                                        Yes
    3                No         No                                        Yes
    4                No         No                                         No
    …                 …         …                                          …
    308847           No         No                                         No
    308848           No         No                                        Yes
    308849           No        Yes  Yes, but female told only during pregnancy
    308850           No         No                                         No
    308851           No         No                                         No

           Arthritis     Sex Age_Category  Height_(cm)  Weight_(kg)    BMI  \
    0            Yes  Female        70-74          150        32.66  14.54
    1             No  Female        70-74          165        77.11  28.29
    2             No  Female        60-64          163        88.45  33.47
    3             No    Male        75-79          180        93.44  28.73
    4             No    Male          80+          191        88.45  24.37
    …              …     …           …            …          …     …
    308847        No    Male        25-29          168        81.65  29.05
    308848        No    Male        65-69          180        69.85  21.48
    308849        No  Female        30-34          157        61.23  24.69
    308850        No    Male        65-69          183        79.38  23.73
    308851        No  Female        45-49          160        81.19  31.71

           Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
    0                  Yes                    0                 30
    1                   No                    0                 30
    2                   No                    4                 12
    3                   No                    0                 30
```

```
4                  Yes                     0                   8
…                  …                       …                   …
308847              No                     4                  30
308848              No                     8                  15
308849             Yes                     4                  40
308850              No                     3                  30
308851              No                     1                   5

        Green_Vegetables_Consumption  FriedPotato_Consumption Heart_Disease
0                                 16                      12            No
1                                  0                       4           Yes
2                                  3                      16            No
3                                 30                       8           Yes
4                                  4                       0            No
…                                  …                       …             …
308847                             8                       0            No
308848                            60                       4            No
308849                             8                       4            No
308850                            12                       0            No
308851                            12                       1            No

[308852 rows x 19 columns]
```

DATA PREPROCESSING

[ ]:
```
#Printing first five values
df.head()
```

[ ]:
```
  General_Health                  Checkup Exercise Skin_Cancer Other_Cancer  \
0           Poor  Within the past 2 years       No          No           No
1      Very Good     Within the past year       No          No           No
2      Very Good     Within the past year      Yes          No           No
3           Poor     Within the past year      Yes          No           No
4           Good     Within the past year       No          No           No

  Depression Diabetes Arthritis     Sex Age_Category  Height_(cm)  \
0         No       No       Yes  Female        70-74          150
1         No      Yes        No  Female        70-74          165
2         No      Yes        No  Female        60-64          163
3         No      Yes        No    Male        75-79          180
4         No       No        No    Male          80+          191

   Weight_(kg)    BMI Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
0        32.66  14.54             Yes                    0                 30
1        77.11  28.29              No                    0                 30
2        88.45  33.47              No                    4                 12
3        93.44  28.73              No                    0                 30
```

```
4        88.45  24.37             Yes                     0                  8

   Green_Vegetables_Consumption  FriedPotato_Consumption Heart_Disease
0                            16                       12            No
1                             0                        4           Yes
2                             3                       16            No
3                            30                        8           Yes
4                             4                        0            No
```

[ ]:
```
#Printing last five values
df.tail()
```

[ ]:
```
       General_Health                  Checkup Exercise Skin_Cancer  \
308847      Very Good     Within the past year      Yes          No
308848           Fair  Within the past 5 years      Yes          No
308849      Very Good      5 or more years ago      Yes          No
308850      Very Good     Within the past year      Yes          No
308851      Excellent     Within the past year      Yes          No

       Other_Cancer Depression                                      Diabetes  \
308847           No         No                                           No
308848           No         No                                          Yes
308849           No        Yes  Yes, but female told only during pregnancy
308850           No         No                                           No
308851           No         No                                           No

       Arthritis     Sex Age_Category  Height_(cm)  Weight_(kg)    BMI  \
308847        No    Male        25-29          168        81.65  29.05
308848        No    Male        65-69          180        69.85  21.48
308849        No  Female        30-34          157        61.23  24.69
308850        No    Male        65-69          183        79.38  23.73
308851        No  Female        45-49          160        81.19  31.71

       Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
308847              No                    4                 30
308848              No                    8                 15
308849             Yes                    4                 40
308850              No                    3                 30
308851              No                    1                  5

        Green_Vegetables_Consumption  FriedPotato_Consumption Heart_Disease
308847                             8                       0            No
308848                            60                       4            No
308849                             8                       4            No
308850                            12                       0            No
308851                            12                       1            No
```

[ ]:
```
#Number of rows and columns
df.shape
```

[ ]:
```
(308852, 19)
```

[ ]:
```
#Printing columns
df.columns
```

[ ]:
```
Index(['General_Health', 'Checkup', 'Exercise', 'Skin_Cancer', 'Other_Cancer',
       'Depression', 'Diabetes', 'Arthritis', 'Sex', 'Age_Category',
       'Height_(cm)', 'Weight_(kg)', 'BMI', 'Smoking_History',
       'Alcohol_Consumption', 'Fruit_Consumption',
       'Green_Vegetables_Consumption', 'FriedPotato_Consumption',
       'Heart_Disease'],
      dtype='object')
```

[ ]:
```
#Printing datatypes
df.dtypes
```

[ ]:
```
General_Health                  object
Checkup                         object
Exercise                        object
Skin_Cancer                     object
Other_Cancer                    object
Depression                      object
Diabetes                        object
Arthritis                       object
Sex                             object
Age_Category                    object
Height_(cm)                      int64
Weight_(kg)                    float64
BMI                            float64
Smoking_History                 object
Alcohol_Consumption              int64
Fruit_Consumption                int64
Green_Vegetables_Consumption     int64
FriedPotato_Consumption          int64
Heart_Disease                   object
dtype: object
```

[ ]:
```
#Unique values
df.nunique()
```

[ ]:
```
General_Health                   5
Checkup                          5
Exercise                         2
Skin_Cancer                      2
```

```
Other_Cancer                      2
Depression                        2
Diabetes                          4
Arthritis                         2
Sex                               2
Age_Category                     13
Height_(cm)                      99
Weight_(kg)                     525
BMI                            3654
Smoking_History                   2
Alcohol_Consumption              31
Fruit_Consumption                77
Green_Vegetables_Consumption     75
FriedPotato_Consumption          69
Heart_Disease                     2
dtype: int64
```

[ ]:
```
df.describe()
#helps us to understand how data has been spread across the table.
#count :- the number of Non-empty rows in each features.
#mean :- mean value of that feature.
#std :- Standard Deviation Value of that feature.
#min :- minimum value of that feature.
#max :- maximum value of that feature.
#25%, 50%, and 75% are the percentile/quartile of each features.
```

[ ]:
```
         Height_(cm)    Weight_(kg)            BMI  Alcohol_Consumption  \
count  308852.000000  308852.000000  308852.000000        308852.000000
mean      170.615259      83.588792      28.626256             5.096373
std        10.658000      21.343187       6.522319             8.199789
min        91.000000      24.950000      12.020000             0.000000
25%       163.000000      68.040000      24.210000             0.000000
50%       170.000000      81.650000      27.440000             1.000000
75%       178.000000      95.250000      31.850000             6.000000
max       241.000000     293.020000      99.330000            30.000000

       Fruit_Consumption  Green_Vegetables_Consumption  \
count      308852.000000                 308852.000000
mean           29.835368                     15.110532
std            24.875727                     14.926243
min             0.000000                      0.000000
25%            12.000000                      4.000000
50%            30.000000                     12.000000
75%            30.000000                     20.000000
max           120.000000                    128.000000

       FriedPotato_Consumption
```

```
count            308852.000000
mean                  6.296624
std                   8.582978
min                   0.000000
25%                   2.000000
50%                   4.000000
75%                   8.000000
max                 128.000000
```

[ ]:
```
# finding out missing values
df.isna().sum()
```

[ ]:
```
General_Health                  0
Checkup                         0
Exercise                        0
Skin_Cancer                     0
Other_Cancer                    0
Depression                      0
Diabetes                        0
Arthritis                       0
Sex                             0
Age_Category                    0
Height_(cm)                     0
Weight_(kg)                     0
BMI                             0
Smoking_History                 0
Alcohol_Consumption             0
Fruit_Consumption               0
Green_Vegetables_Consumption    0
FriedPotato_Consumption         0
Heart_Disease                   0
dtype: int64
```

[ ]:
```
# Removing duplicate rows
df.drop_duplicates(ignore_index=True)
```

[ ]:
```
       General_Health                  Checkup Exercise Skin_Cancer  \
0                Poor  Within the past 2 years       No          No
1           Very Good     Within the past year       No          No
2           Very Good     Within the past year      Yes          No
3                Poor     Within the past year      Yes          No
4                Good     Within the past year       No          No
…                 …                        …      …           …
308767      Very Good     Within the past year      Yes          No
308768           Fair  Within the past 5 years      Yes          No
308769      Very Good      5 or more years ago      Yes          No
308770      Very Good     Within the past year      Yes          No
```

```
308771      Excellent     Within the past year      Yes          No

       Other_Cancer Depression                                    Diabetes  \
0                No         No                                         No
1                No         No                                        Yes
2                No         No                                        Yes
3                No         No                                        Yes
4                No         No                                         No
…              …          …                                        …
308767           No         No                                         No
308768           No         No                                        Yes
308769           No        Yes  Yes, but female told only during pregnancy
308770           No         No                                         No
308771           No         No                                         No

       Arthritis     Sex Age_Category  Height_(cm)  Weight_(kg)    BMI  \
0            Yes  Female        70-74          150        32.66  14.54
1             No  Female        70-74          165        77.11  28.29
2             No  Female        60-64          163        88.45  33.47
3             No    Male        75-79          180        93.44  28.73
4             No    Male          80+          191        88.45  24.37
…           …     …           …          …           …    …
308767        No    Male        25-29          168        81.65  29.05
308768        No    Male        65-69          180        69.85  21.48
308769        No  Female        30-34          157        61.23  24.69
308770        No    Male        65-69          183        79.38  23.73
308771        No  Female        45-49          160        81.19  31.71

       Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
0                  Yes                    0                 30
1                   No                    0                 30
2                   No                    4                 12
3                   No                    0                 30
4                  Yes                    0                  8
…                …                      …                …
308767              No                    4                 30
308768              No                    8                 15
308769             Yes                    4                 40
308770              No                    3                 30
308771              No                    1                  5

        Green_Vegetables_Consumption  FriedPotato_Consumption Heart_Disease
0                                 16                       12           No
1                                  0                        4          Yes
2                                  3                       16           No
3                                 30                        8          Yes
4                                  4                        0           No
```

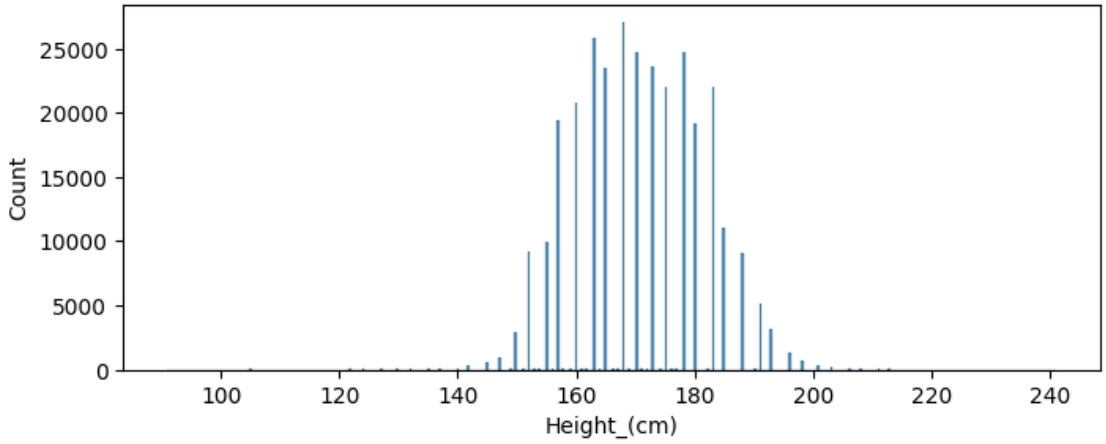
…	…	…	…
308767	8	0	No
308768	60	4	No
308769	8	4	No
308770	12	0	No
308771	12	1	No

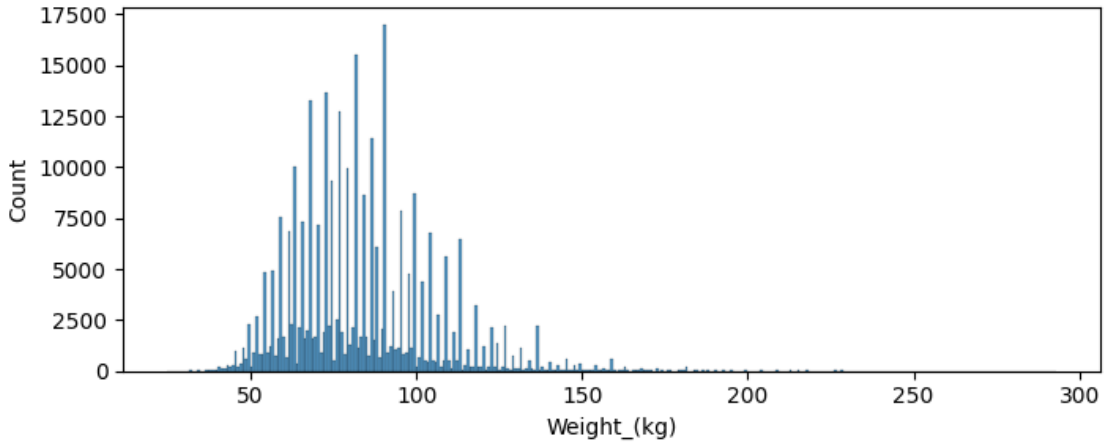
[308772 rows x 19 columns]

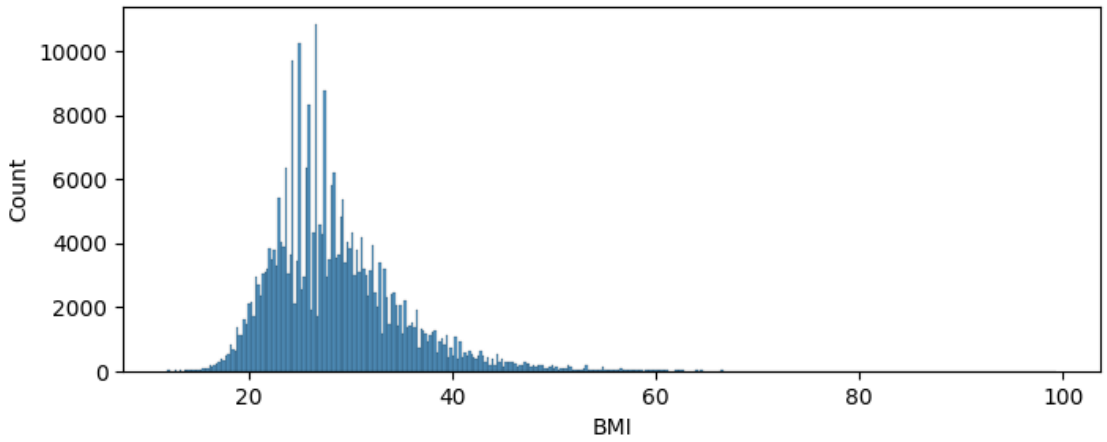
DATA VISUALIZATION

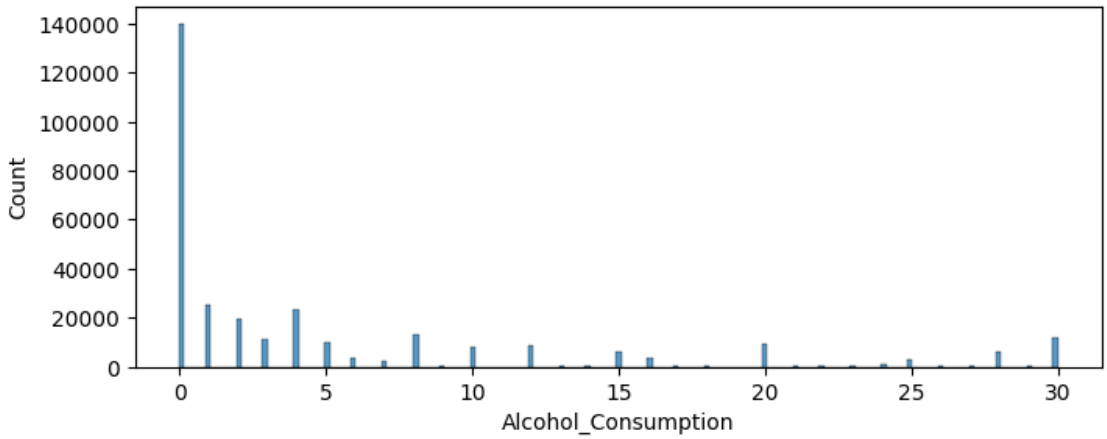
UNIVARIATE ANALYSIS

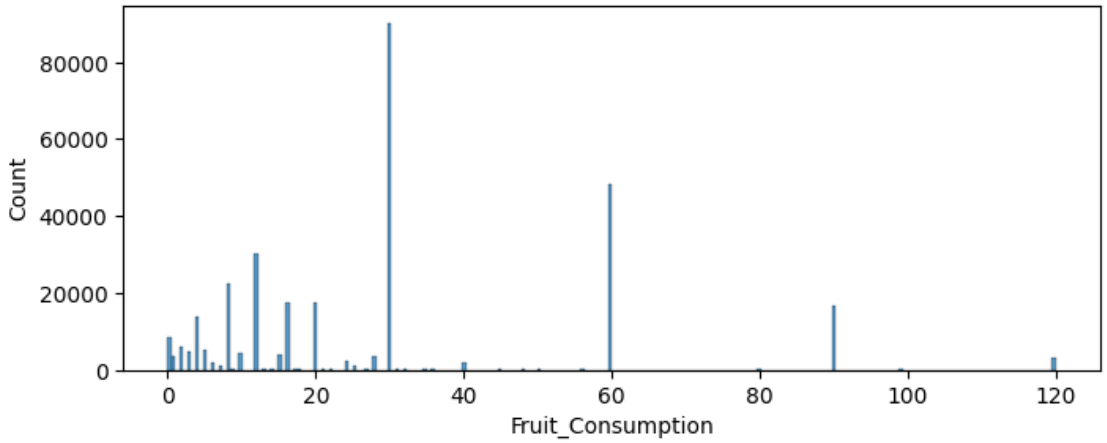
[ ]:
```
# Check the distribution of numerical features
features1=['Height_(cm)','Weight_(kg)','BMI','Alcohol_Consumption','Fruit_Consumption','Green_
for i in features1:
  plt.figure(figsize=(8,3))
  sns.histplot(x=i,data=df)
```

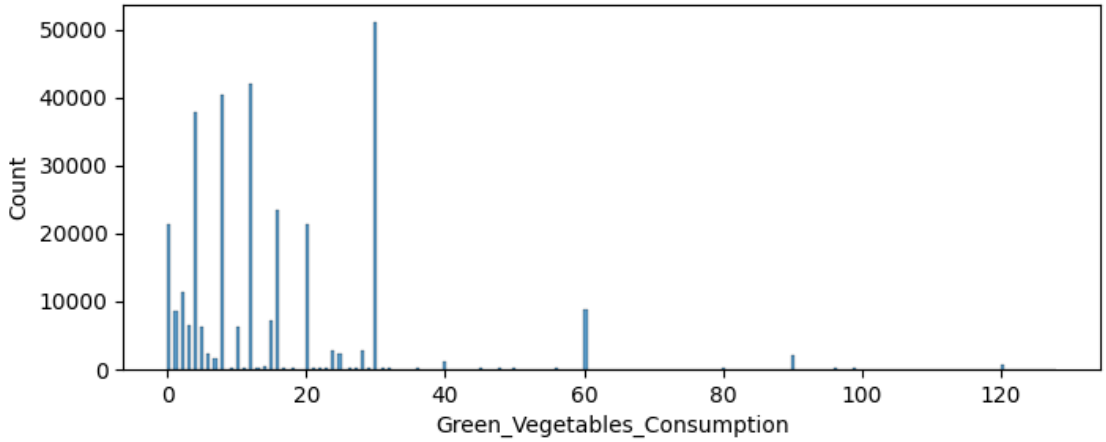


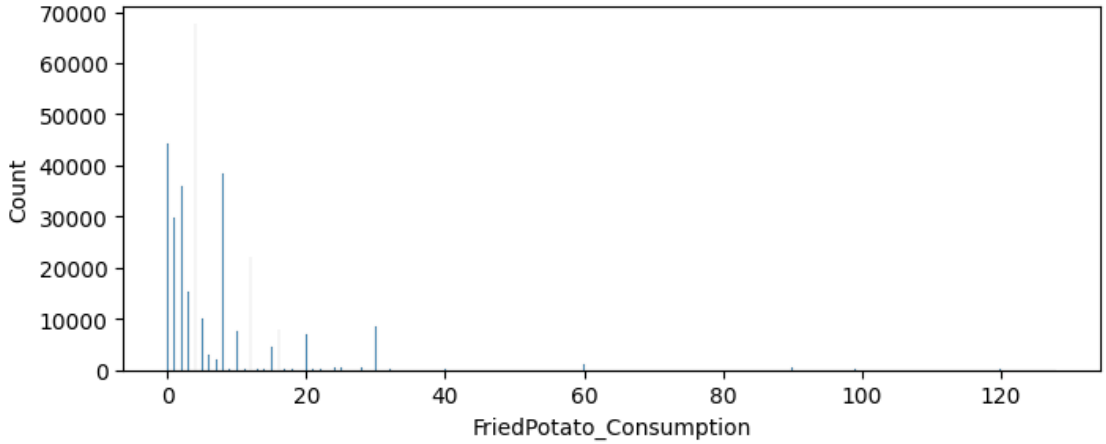



Count
10000
8000
6000
4000
2000
0
20
40
60
80
100
BMI


Count
140000
120000
100000
80000
60000
40000
20000
0
0
5
10
15
20
25
30
Alcohol_Consumption


Count
80000
60000
40000
20000
0
0
20
40
60
80
100
120
Fruit_Consumption

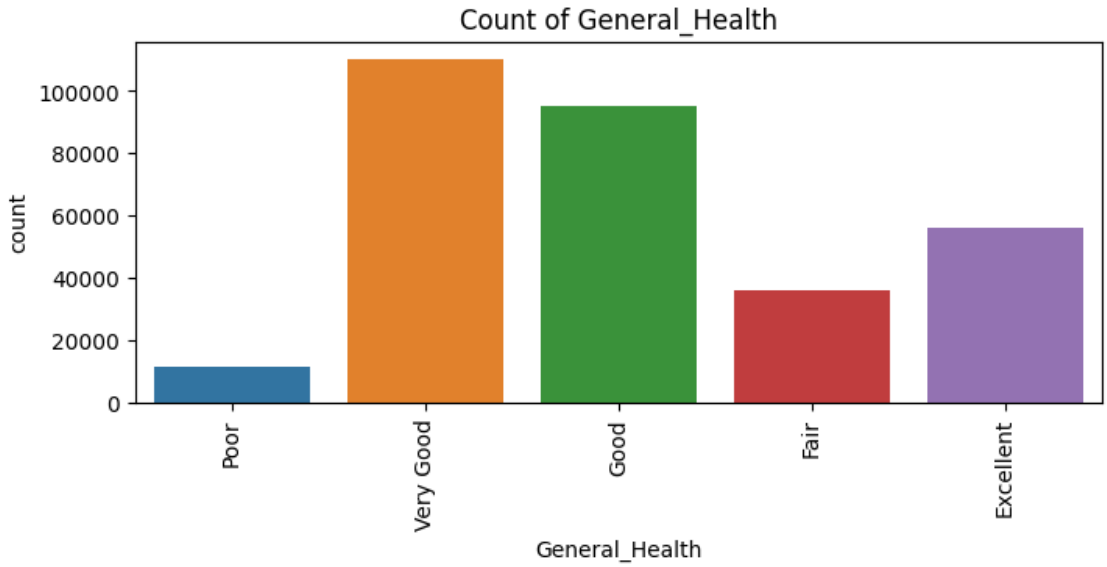


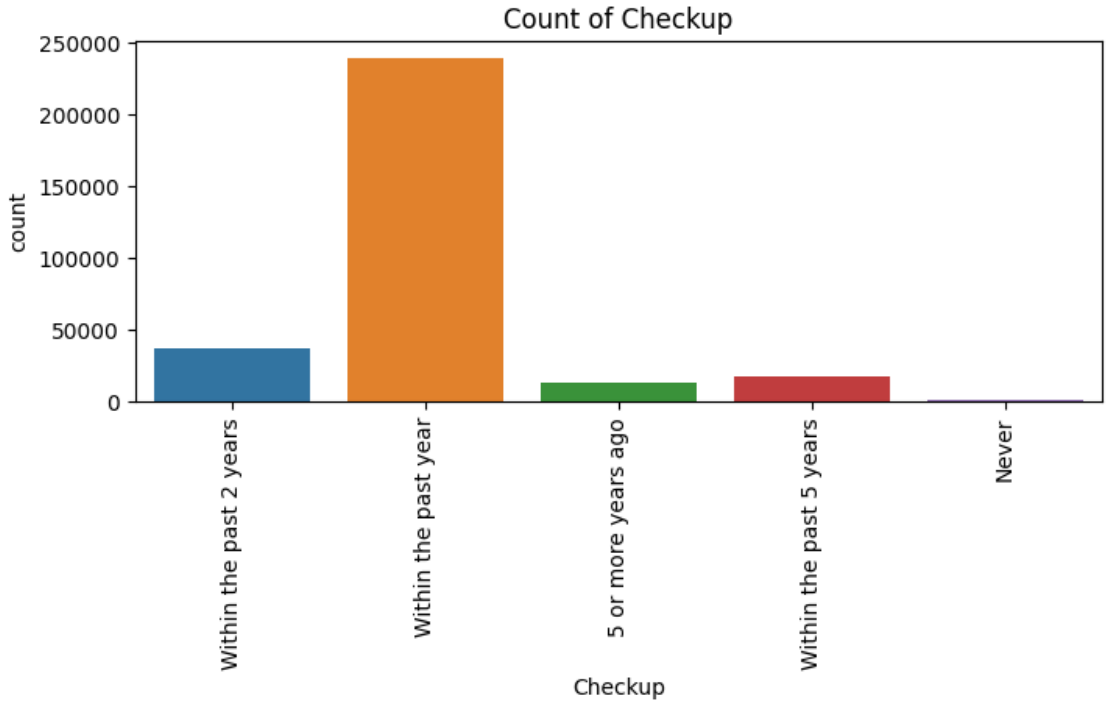


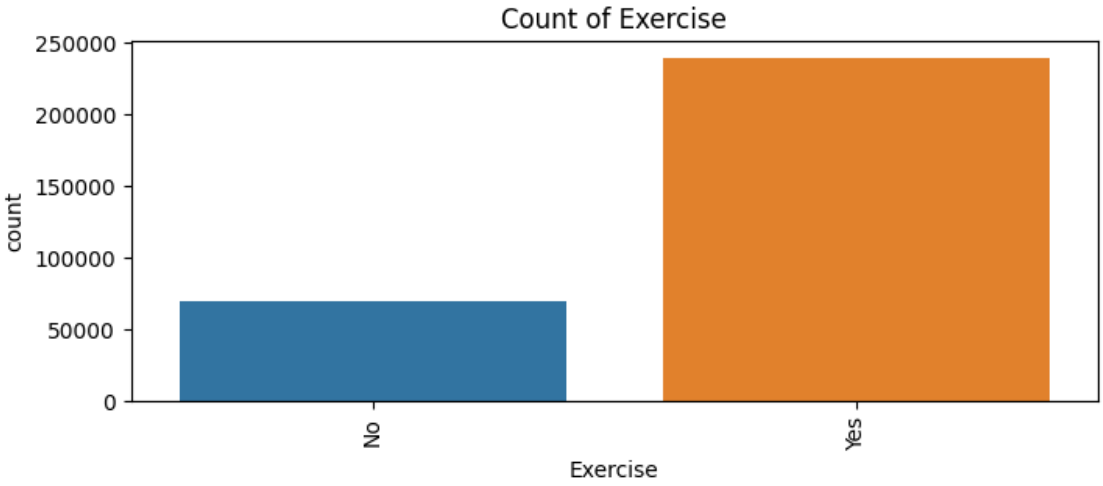
OBSERVATION

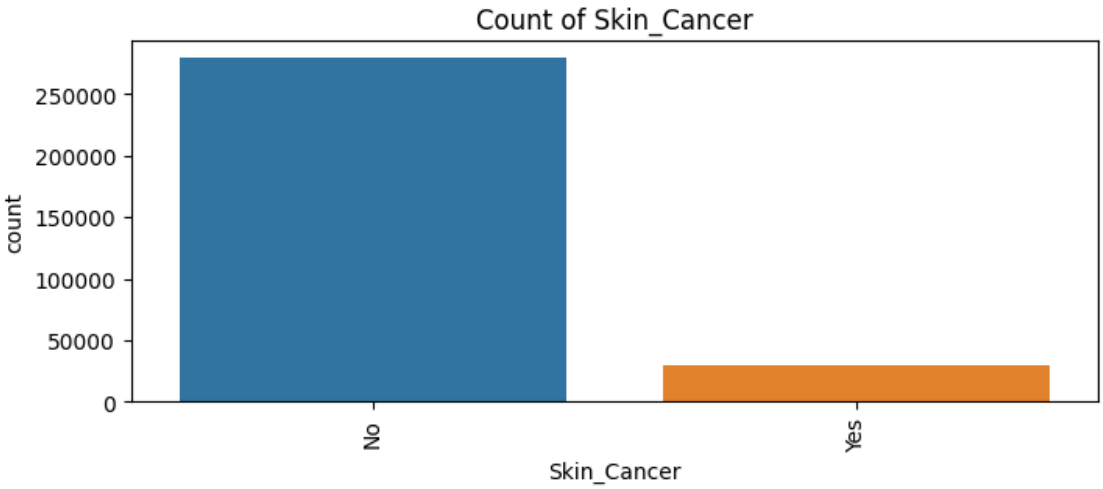
1. Height_(cm): The distribution appears to be normal with a peak around 170 cm.
2. Weight_(kg): The distribution is slightly right-skewed with a peak around 80 kg.
3. BMI: The distribution is slightly right-skewed with a peak around 25-30, which is the normal range of BMI.
4. Alcohol_Consumption: Most of the participants consume little to no alcohol, as the distribution is heavily right-skewed.
5. Fruit_Consumption: A significant number of individuals consume around 30 units, but there's also a large number who consume very little.
6. Green_Vegetables_Consumption: The consumption of green vegetables is spread quite evenly, with slightly more people consuming very little.
7. FriedPotato_Consumption: Most of the participants consume little to no fried potatoes, as the distribution is heavily right-skewed.

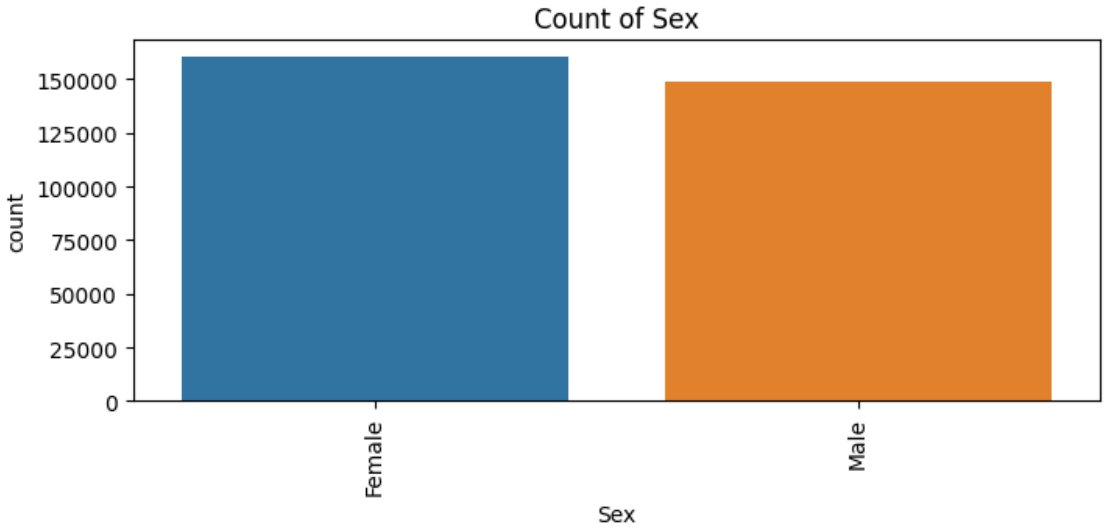
[ ]:
```
# check the distribution of categorical features
features2=['General_Health','Checkup','Exercise','Skin_Cancer','Sex','Other_Cancer','Depressio
for i in features2:
  plt.figure(figsize=(8,3))
  sns.countplot(x=i,data=df,hue=i)
  plt.xticks(rotation=90)
  plt.title('Count of ' + i)
```

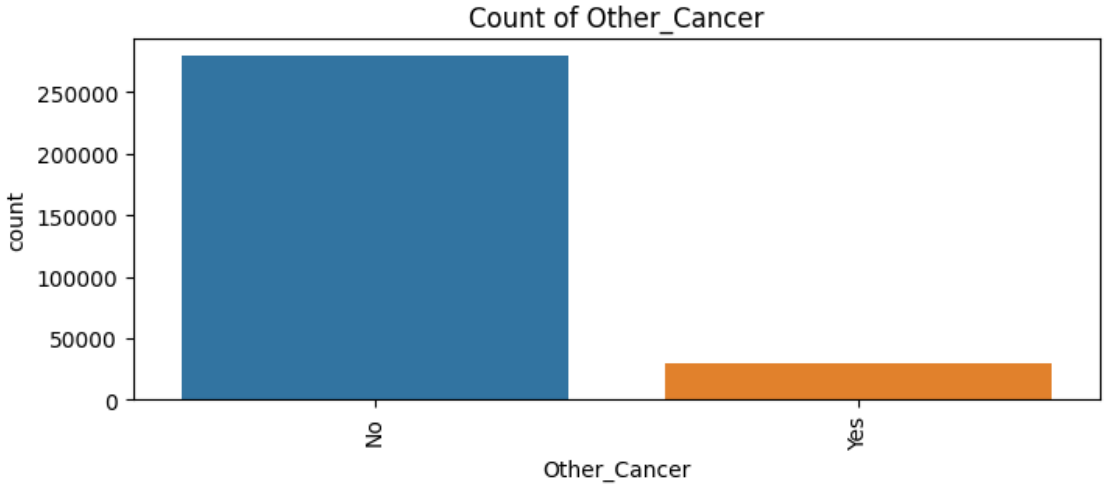


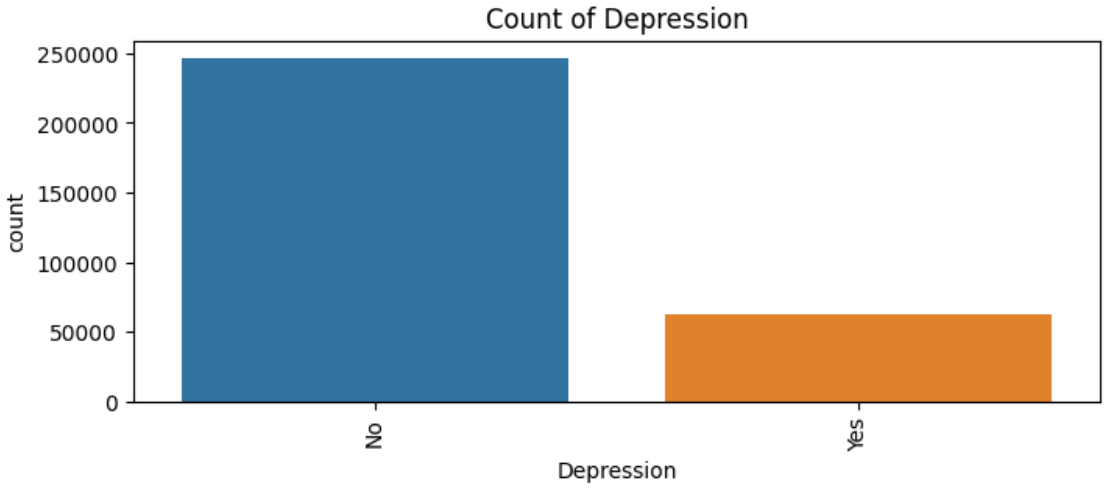


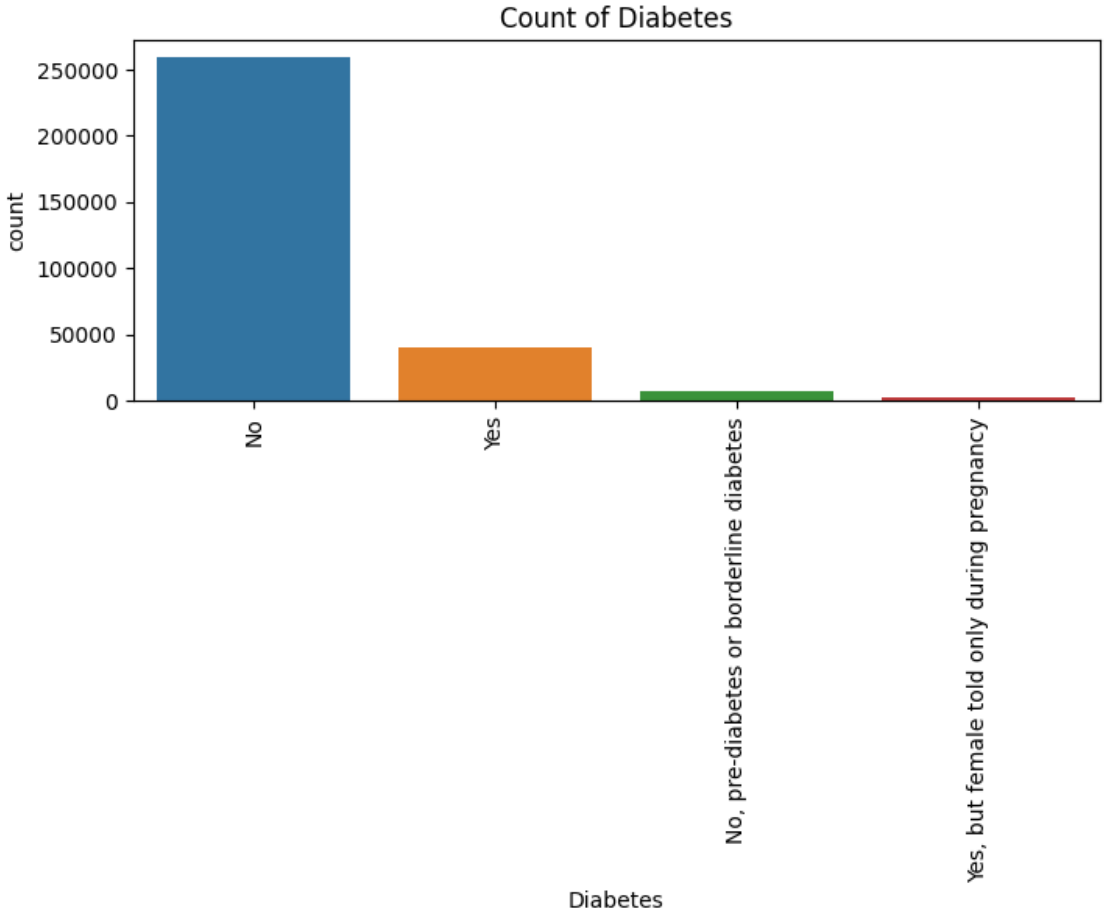

Count of Exercise
250000
200000
150000
100000
50000
0
count
No
Yes
Exercise

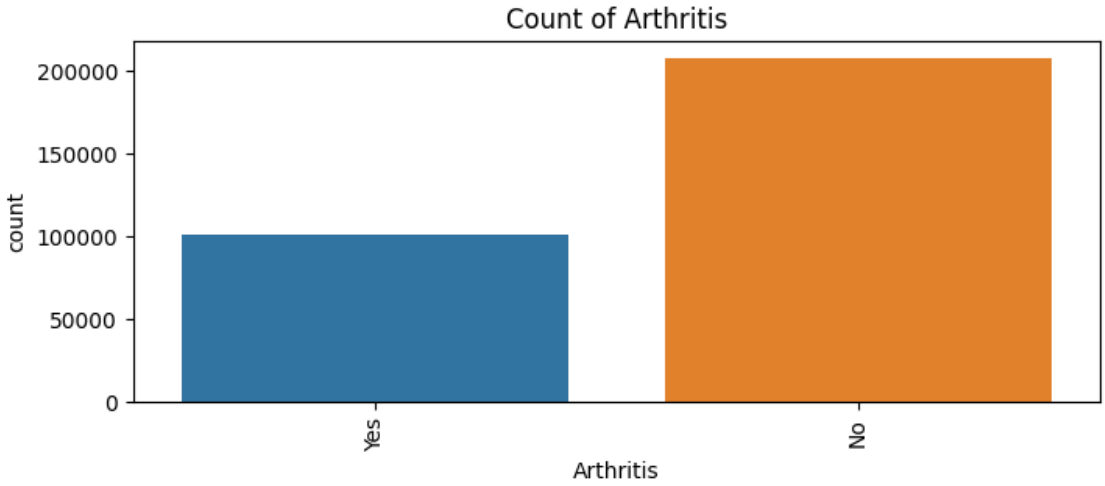

Count of Skin_Cancer
250000
200000
150000
100000
50000
0
count
No
Yes
Skin_Cancer

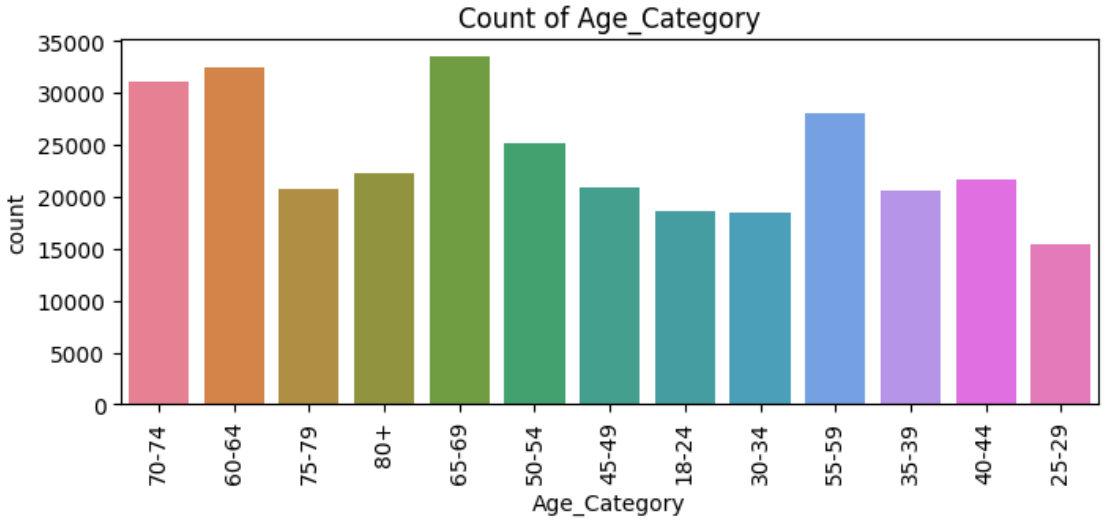

Count of Sex
150000
125000
100000
75000
50000
25000
0
count
Female
Male
Sex

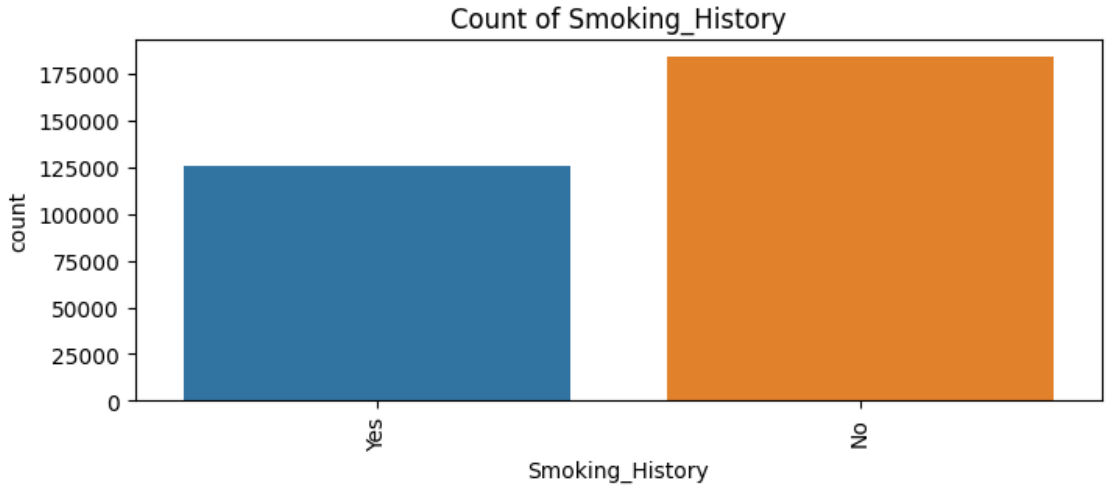

Count of Other_Cancer
250000
200000
150000
100000
50000
0
count
No
Yes
Other_Cancer

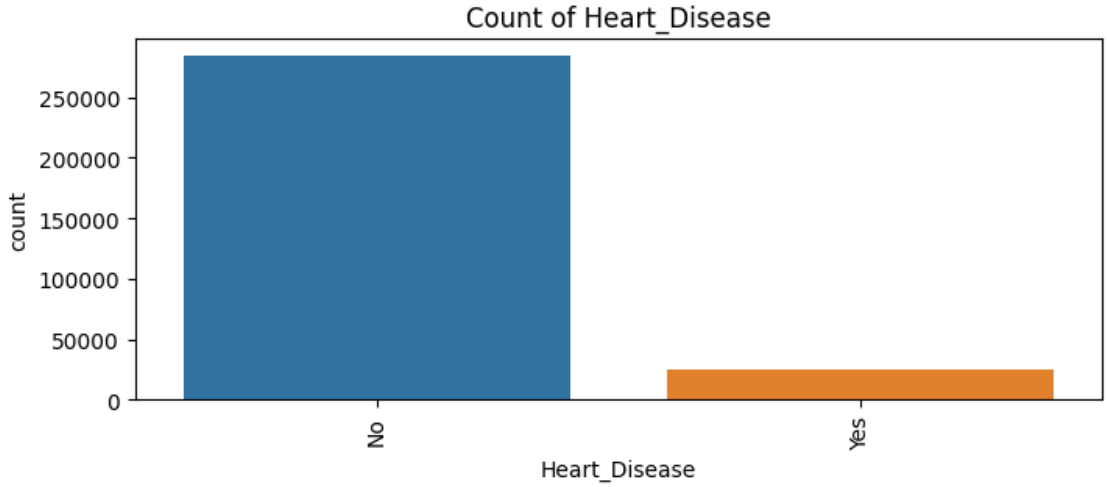

Count of Depression
250000
200000
150000
100000
50000
0
count
No
Yes
Depression


Count of Diabetes
250000
200000
150000
100000
50000
0
count
No
Yes
No, pre-diabetes or borderline diabetes
Yes, but female told only during pregnancy
Diabetes


Count of Arthritis
200000
150000
100000
50000
0
count
Yes
No
Arthritis


Count of Age_Category
35000
30000
25000
20000
15000
10000
5000
0
count
70-74
60-64
75-79
80+
65-69
50-54
45-49
18-24
30-34
55-59
35-39
40-44
25-29
Age_Category





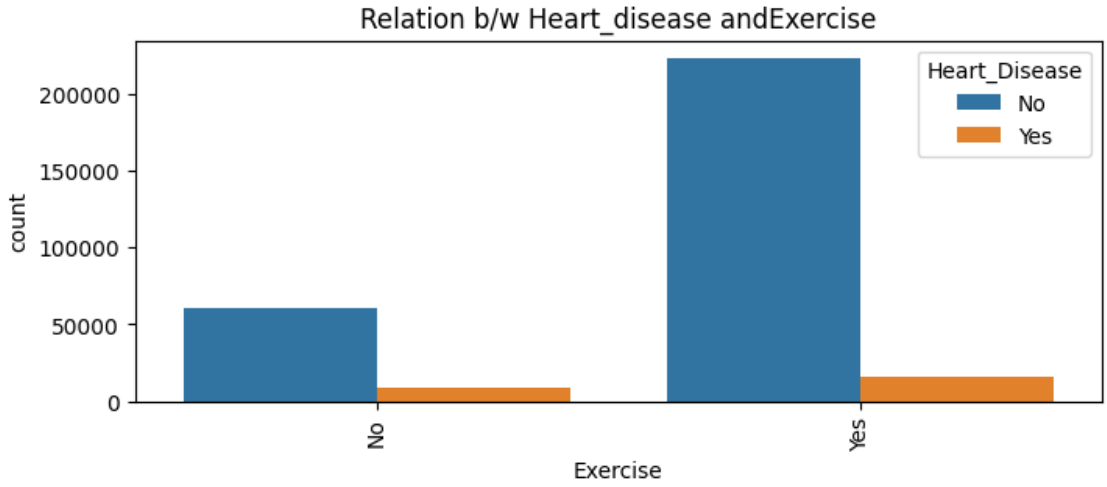
OBSERVATIONS FROM COUNTPLOT

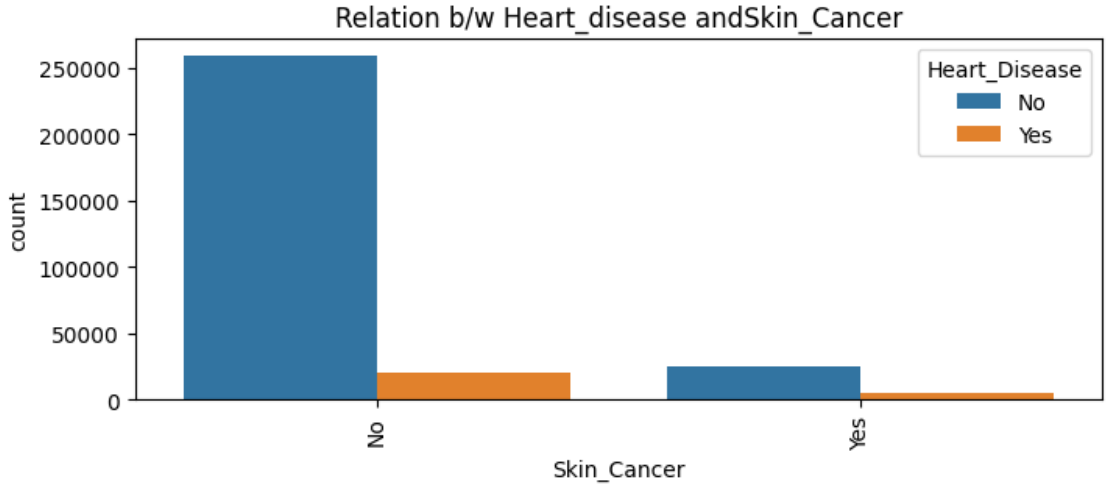
1. General_Health: Most individuals report their general health as "Very Good". "Poor" health is the least reported.
2. Checkup: The majority of individuals have had a checkup within the past year.
3. Exercise: Most individuals report that they exercise.
4. Skin_Cancer: A vast majority of individuals do not have skin cancer.
5. Sex: There are slightly more females than males in the dataset.
6. Other_Cancer: Similar to skin cancer, most individuals do not have other types of cancer.
7. Depression: Most individuals do not have depression.
8. Diabetes: The majority of individuals do not have diabetes.
9. Arthritis: More individuals do not have arthritis than those who do.
10. Smoking_History: Most individuals do not have a smoking history.
11. Age_Category: The age categories are quite evenly distributed, with a slightly higher count

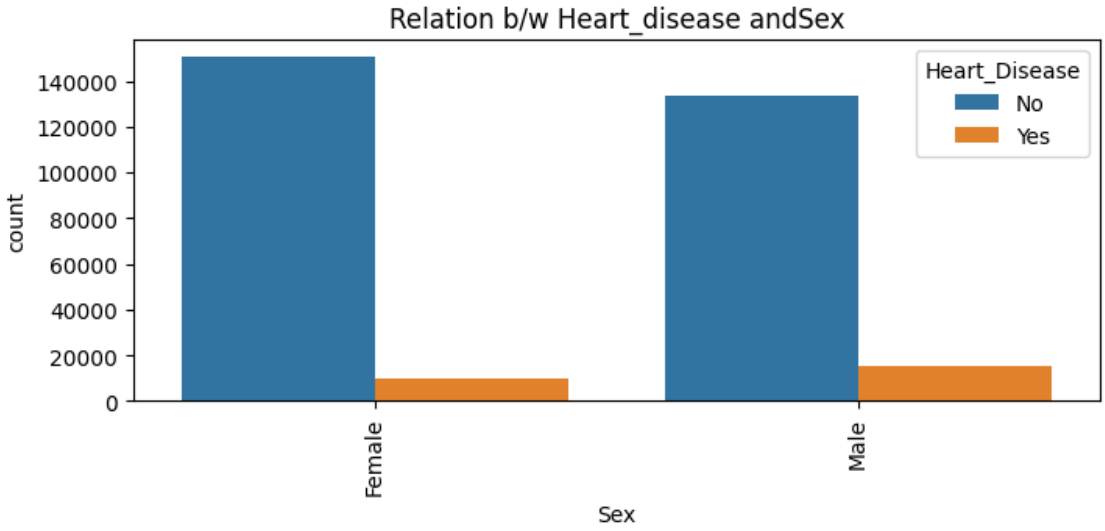
in the “65-69” category.

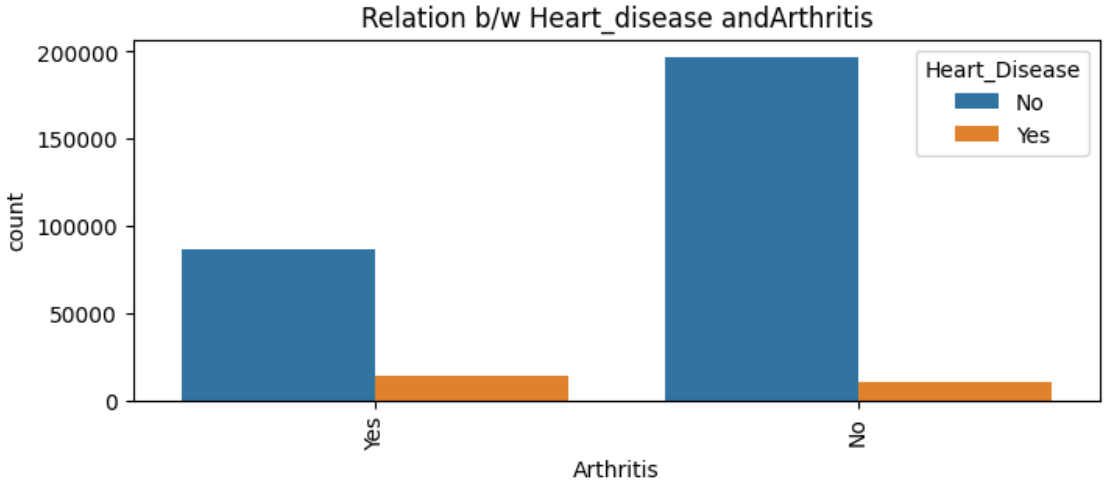
12. Heart_Disease: The majority of individuals do not have heart disease.

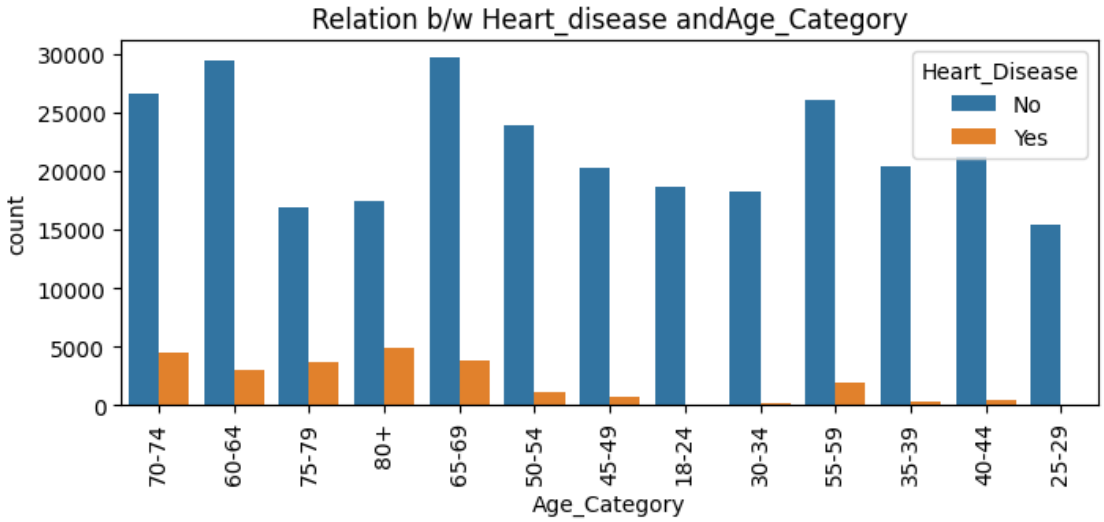
```
[ ]: # analysing the given features and heart disease
features3=['Exercise','Skin_Cancer','Sex','Arthritis','Age_Category','Smoking_History']
for i in features3:
  plt.figure(figsize=(8,3))
  sns.countplot(x=i,data=df,hue='Heart_Disease')
  plt.xticks(rotation=90)
  plt.title('Relation b/w Heart_disease and' + i)
```

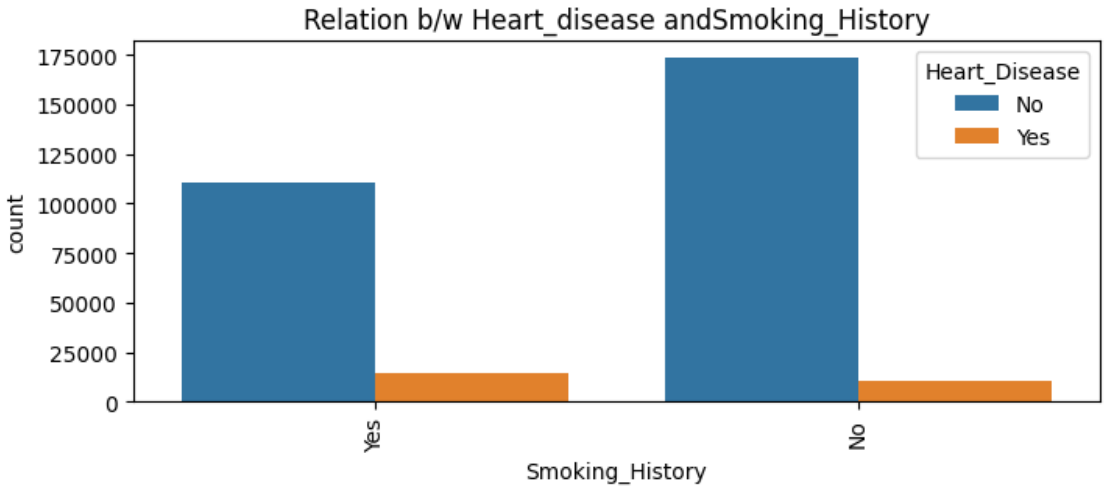





Relation b/w Heart_disease andSex
Heart_Disease
No
Yes
count
140000
120000
100000
80000
60000
40000
20000
0
Female
Male
Sex


Relation b/w Heart_disease andArthritis
Heart_Disease
No
Yes
count
200000
150000
100000
50000
0
Yes
No
Arthritis



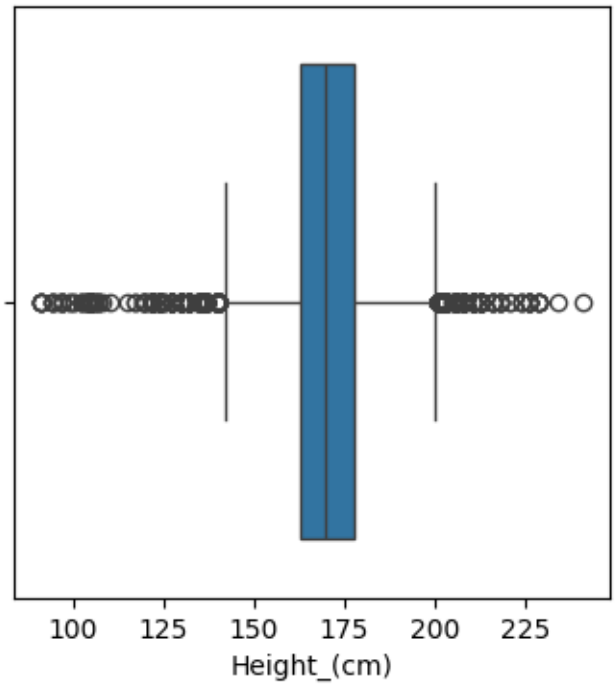


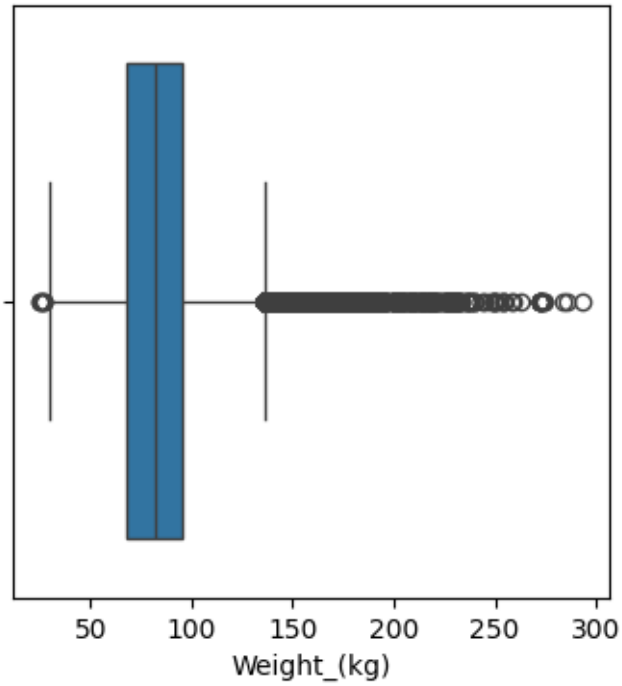
OBSERVATION FROM THE ABOVE ANLAYSIS

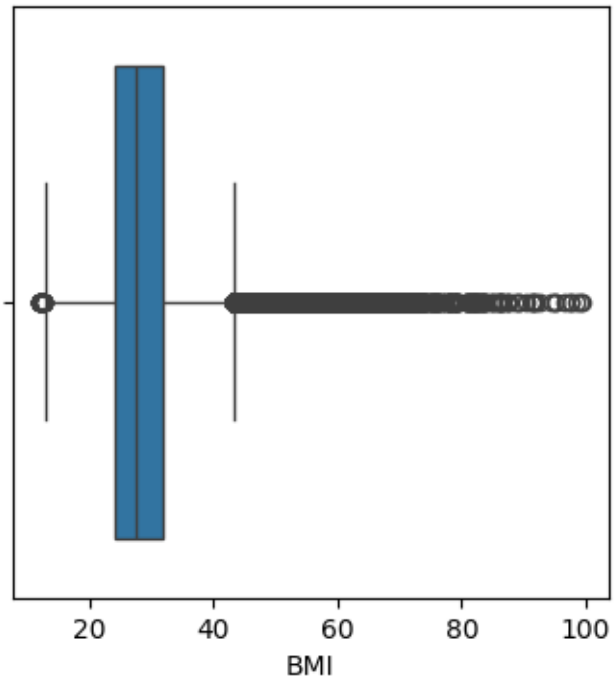
- The person who having arthritis have more chances for heart disease.
- It is slightly more common in patients who do not exercise.
- Males are more likely to have heart disease than females
- The prevalence of heart disease increases with age, with it being most common in the 80+ age category.
- Heart disease is also more common in patients with a history of smoking

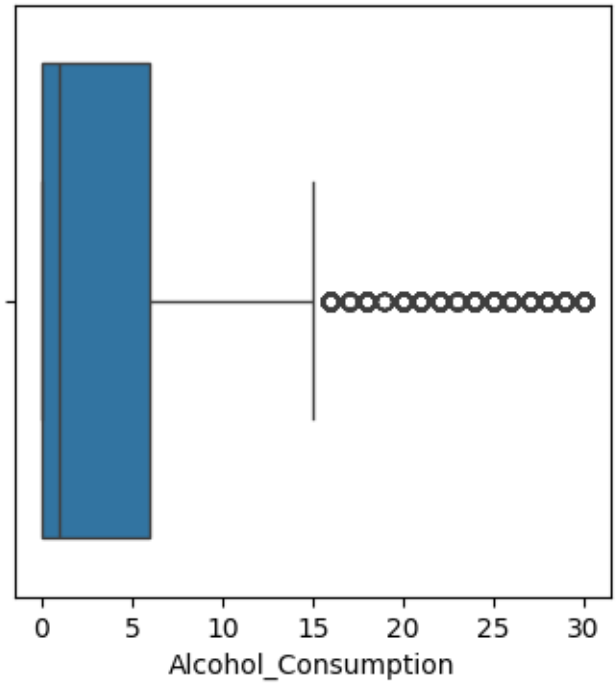
ANALYSING THE OUTLIERS

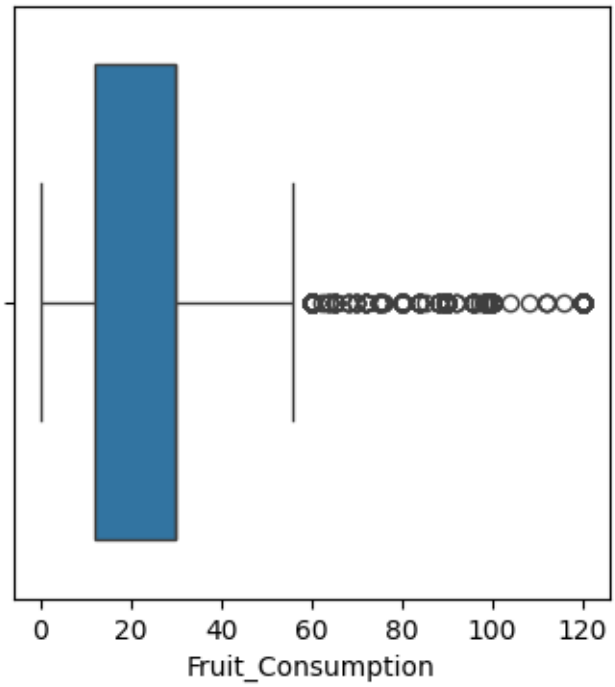
```
[ ]: outlier_features=['Height_(cm)','Weight_(kg)','BMI','Alcohol_Consumption','Fruit_Consumption',
     for i in outlier_features:
         plt.figure(figsize=(4,4))
         sns.boxplot(x=i,data=df)
```

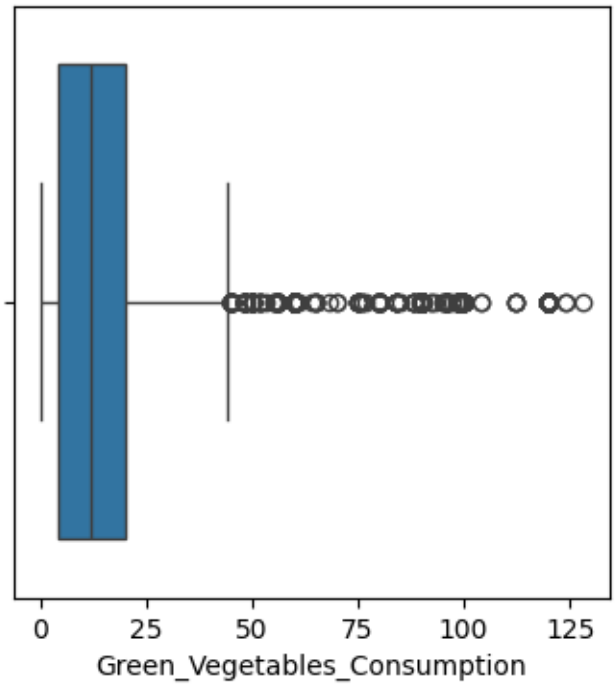


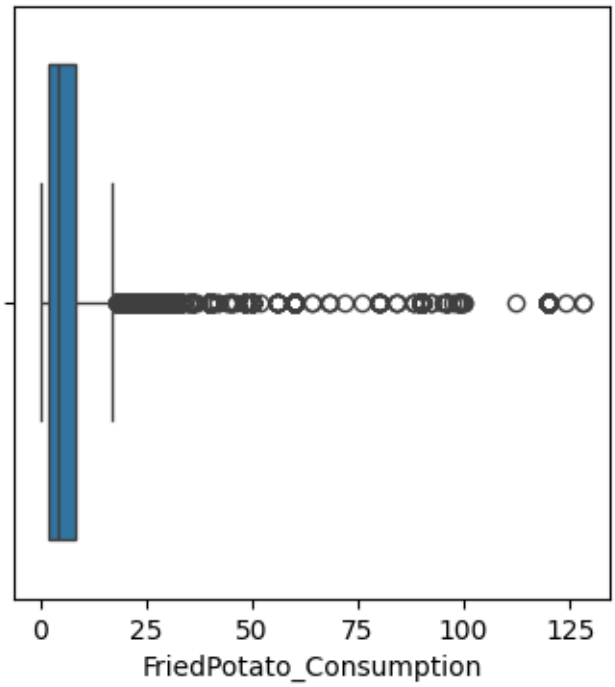



20
40
60
80
100
BMI


0
5
10
15
20
25
30
Alcohol_Consumption


0
20
40
60
80
100
120
Fruit_Consumption


0
25
50
75
100
125
Green_Vegetables_Consumption



INTER QUARTILE RANGE METHOD

[ ]:
```
for i in␣
 ↪['Height_(cm)','Weight_(kg)','BMI','Alcohol_Consumption','Fruit_Consumption','Green_Vegetab
 ↪
  Q1=df[i].quantile(0.25)
  Q3=df[i].quantile(0.75)
  IQR=Q3-Q1
  UPPER=Q3+(1.5*IQR)
  LOWER=Q1-(1.5*IQR)
  dfb=df[(df[i]>=LOWER)&(df[i]<=UPPER)]
```

[ ]:
```
#Correcting the order of index after removing the outliers
df1=dfb.reset_index(drop=True)
df1
```

[ ]:

	General_Health	Checkup	Exercise	Skin_Cancer	\
0	Poor	Within the past 2 years	No	No	
1	Very Good	Within the past year	No	No	
2	Very Good	Within the past year	Yes	No	
3	Poor	Within the past year	Yes	No	
4	Good	Within the past year	No	No	
…	…	…	…	…	
289391	Very Good	Within the past year	Yes	No	

```
289392           Fair  Within the past 5 years      Yes          No
289393      Very Good      5 or more years ago      Yes          No
289394      Very Good     Within the past year      Yes          No
289395      Excellent     Within the past year      Yes          No

       Other_Cancer Depression                                      Diabetes  \
0                No         No                                           No
1                No         No                                          Yes
2                No         No                                          Yes
3                No         No                                          Yes
4                No         No                                           No
…              …        …                                          …
289391           No         No                                           No
289392           No         No                                          Yes
289393           No        Yes  Yes, but female told only during pregnancy
289394           No         No                                           No
289395           No         No                                           No

       Arthritis     Sex Age_Category  Height_(cm)  Weight_(kg)    BMI  \
0            Yes  Female        70-74          150        32.66  14.54
1             No  Female        70-74          165        77.11  28.29
2             No  Female        60-64          163        88.45  33.47
3             No    Male        75-79          180        93.44  28.73
4             No    Male          80+          191        88.45  24.37
…          …     …           …           …          …    …
289391        No    Male        25-29          168        81.65  29.05
289392        No    Male        65-69          180        69.85  21.48
289393        No  Female        30-34          157        61.23  24.69
289394        No    Male        65-69          183        79.38  23.73
289395        No  Female        45-49          160        81.19  31.71

       Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
0                  Yes                    0                 30
1                   No                    0                 30
2                   No                    4                 12
3                   No                    0                 30
4                  Yes                    0                  8
…                …                      …                …
289391              No                    4                 30
289392              No                    8                 15
289393             Yes                    4                 40
289394              No                    3                 30
289395              No                    1                  5

        Green_Vegetables_Consumption  FriedPotato_Consumption Heart_Disease
0                                 16                       12            No
1                                  0                        4           Yes
```

```
2                                3                     16            No
3                               30                      8           Yes
4                                4                      0            No
…                              …                      …             …
289391                           8                      0            No
289392                          60                      4            No
289393                           8                      4            No
289394                          12                      0            No
289395                          12                      1            No

[289396 rows x 19 columns]
```

ENCODING

[ ]:
```
df1.dtypes
```

[ ]:
```
General_Health                 object
Checkup                        object
Exercise                       object
Skin_Cancer                    object
Other_Cancer                   object
Depression                     object
Diabetes                       object
Arthritis                      object
Sex                            object
Age_Category                   object
Height_(cm)                     int64
Weight_(kg)                   float64
BMI                           float64
Smoking_History                object
Alcohol_Consumption             int64
Fruit_Consumption               int64
Green_Vegetables_Consumption    int64
FriedPotato_Consumption         int64
Heart_Disease                  object
dtype: object
```

[ ]:
```
# binary columns having yes or no values encoding using Label encoder
binary_columns=['Exercise','Skin_Cancer','Other_Cancer','Depression','Arthritis','Smoking_Hist
from sklearn.preprocessing import LabelEncoder
end=LabelEncoder()
for i in binary_columns:
    df1[i]=end.fit_transform(df1[i])
df1
```

[ ]:
```
       General_Health                 Checkup  Exercise  Skin_Cancer  \
0                Poor  Within the past 2 years         0            0
```

```
1           Very Good     Within the past year         0            0
2           Very Good     Within the past year         1            0
3                Poor     Within the past year         1            0
4                Good     Within the past year         0            0
…                 …                        …       …          …
289391      Very Good     Within the past year         1            0
289392           Fair  Within the past 5 years         1            0
289393      Very Good      5 or more years ago         1            0
289394      Very Good     Within the past year         1            0
289395      Excellent     Within the past year         1            0

        Other_Cancer  Depression                                    Diabetes  \
0                  0           0                                         No
1                  0           0                                        Yes
2                  0           0                                        Yes
3                  0           0                                        Yes
4                  0           0                                         No
…                …           …                                    …
289391             0           0                                         No
289392             0           0                                        Yes
289393             0           1  Yes, but female told only during pregnancy
289394             0           0                                         No
289395             0           0                                         No

        Arthritis     Sex Age_Category  Height_(cm)  Weight_(kg)    BMI  \
0               1  Female        70-74          150        32.66  14.54
1               0  Female        70-74          165        77.11  28.29
2               0  Female        60-64          163        88.45  33.47
3               0    Male        75-79          180        93.44  28.73
4               0    Male          80+          191        88.45  24.37
…             …      …           …            …          …    …
289391          0    Male        25-29          168        81.65  29.05
289392          0    Male        65-69          180        69.85  21.48
289393          0  Female        30-34          157        61.23  24.69
289394          0    Male        65-69          183        79.38  23.73
289395          0  Female        45-49          160        81.19  31.71

        Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
0                     1                    0                 30
1                     0                    0                 30
2                     0                    4                 12
3                     0                    0                 30
4                     1                    0                  8
…                   …                    …                 …
289391                0                    4                 30
289392                0                    8                 15
289393                1                    4                 40
```

```
289394                0                   3                 30
289395                0                   1                  5

        Green_Vegetables_Consumption  FriedPotato_Consumption  Heart_Disease
0                                 16                       12              0
1                                  0                        4              1
2                                  3                       16              0
3                                 30                        8              1
4                                  4                        0              0
…                                  …                        …              …
289391                             8                        0              0
289392                            60                        4              0
289393                             8                        4              0
289394                            12                        0              0
289395                            12                        1              0

[289396 rows x 19 columns]
```

[ ]:
```
# Encoding sec category using get dummies
df2=pd.get_dummies(df1[['Sex']])
df2
```

[ ]:
```
        Sex_Female  Sex_Male
0                1         0
1                1         0
2                1         0
3                0         1
4                0         1
…              …         …
289391           0         1
289392           0         1
289393           1         0
289394           0         1
289395           1         0

[289396 rows x 2 columns]
```

[ ]:
```
# Combining two dataframes
dfe=pd.concat([df1,df2],axis=1)
dfe
```

[ ]:
```
       General_Health                  Checkup  Exercise  Skin_Cancer  \
0                Poor  Within the past 2 years         0            0
1           Very Good     Within the past year         0            0
2           Very Good     Within the past year         1            0
3                Poor     Within the past year         1            0
4                Good     Within the past year         0            0
```

```
…              …                          …       …           …
289391      Very Good     Within the past year         1           0
289392           Fair  Within the past 5 years         1           0
289393      Very Good      5 or more years ago         1           0
289394      Very Good     Within the past year         1           0
289395      Excellent     Within the past year         1           0

        Other_Cancer  Depression                                    Diabetes  \
0                  0           0                                         No
1                  0           0                                        Yes
2                  0           0                                        Yes
3                  0           0                                        Yes
4                  0           0                                         No
…              …           …                                    …
289391             0           0                                         No
289392             0           0                                        Yes
289393             0           1  Yes, but female told only during pregnancy
289394             0           0                                         No
289395             0           0                                         No

        Arthritis     Sex Age_Category  …  Weight_(kg)    BMI  \
0               1  Female        70-74  …        32.66  14.54
1               0  Female        70-74  …        77.11  28.29
2               0  Female        60-64  …        88.45  33.47
3               0    Male        75-79  …        93.44  28.73
4               0    Male          80+  …        88.45  24.37
…           …      …           …  …           …     …
289391          0    Male        25-29  …        81.65  29.05
289392          0    Male        65-69  …        69.85  21.48
289393          0  Female        30-34  …        61.23  24.69
289394          0    Male        65-69  …        79.38  23.73
289395          0  Female        45-49  …        81.19  31.71

        Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
0                     1                    0                 30
1                     0                    0                 30
2                     0                    4                 12
3                     0                    0                 30
4                     1                    0                  8
…                 …                    …                 …
289391                0                    4                 30
289392                0                    8                 15
289393                1                    4                 40
289394                0                    3                 30
289395                0                    1                  5

        Green_Vegetables_Consumption  FriedPotato_Consumption  Heart_Disease  \
```

```
0                                 16                      12              0
1                                  0                       4              1
2                                  3                      16              0
3                                 30                       8              1
4                                  4                       0              0
…                                  …                       …              …
289391                             8                       0              0
289392                            60                       4              0
289393                             8                       4              0
289394                            12                       0              0
289395                            12                       1              0

        Sex_Female  Sex_Male
0                1         0
1                1         0
2                1         0
3                0         1
4                0         1
…              …         …
289391           0         1
289392           0         1
289393           1         0
289394           0         1
289395           1         0

[289396 rows x 21 columns]
```

[ ]:
```
dfe.drop('Sex',axis=1,inplace=True)
dfe
```

[ ]:
```
       General_Health                  Checkup  Exercise  Skin_Cancer  \
0                Poor  Within the past 2 years         0            0
1           Very Good     Within the past year         0            0
2           Very Good     Within the past year         1            0
3                Poor     Within the past year         1            0
4                Good     Within the past year         0            0
…                 …                        …        …            …
289391      Very Good     Within the past year         1            0
289392           Fair  Within the past 5 years         1            0
289393      Very Good      5 or more years ago         1            0
289394      Very Good     Within the past year         1            0
289395      Excellent     Within the past year         1            0

        Other_Cancer  Depression                                  Diabetes  \
0                  0           0                                        No
1                  0           0                                       Yes
2                  0           0                                       Yes
```

```
3                 0           0                                          Yes
4                 0           0                                           No
…               …           …                                          …
289391            0           0                                           No
289392            0           0                                          Yes
289393            0           1  Yes, but female told only during pregnancy
289394            0           0                                           No
289395            0           0                                           No

        Arthritis Age_Category  Height_(cm)  Weight_(kg)    BMI  \
0               1        70-74          150        32.66  14.54
1               0        70-74          165        77.11  28.29
2               0        60-64          163        88.45  33.47
3               0        75-79          180        93.44  28.73
4               0          80+          191        88.45  24.37
…             …          …            …          …     …
289391          0        25-29          168        81.65  29.05
289392          0        65-69          180        69.85  21.48
289393          0        30-34          157        61.23  24.69
289394          0        65-69          183        79.38  23.73
289395          0        45-49          160        81.19  31.71

        Smoking_History  Alcohol_Consumption  Fruit_Consumption  \
0                     1                    0                 30
1                     0                    0                 30
2                     0                    4                 12
3                     0                    0                 30
4                     1                    0                  8
…                   …                    …                 …
289391                0                    4                 30
289392                0                    8                 15
289393                1                    4                 40
289394                0                    3                 30
289395                0                    1                  5

        Green_Vegetables_Consumption  FriedPotato_Consumption  Heart_Disease  \
0                                 16                       12              0
1                                  0                        4              1
2                                  3                       16              0
3                                 30                        8              1
4                                  4                        0              0
…                                …                        …              …
289391                             8                        0              0
289392                            60                        4              0
289393                             8                        4              0
289394                            12                        0              0
289395                            12                        1              0
```

```
        Sex_Female  Sex_Male
0                1         0
1                1         0
2                1         0
3                0         1
4                0         1
…              …         …
289391           0         1
289392           0         1
289393           1         0
289394           0         1
289395           1         0

[289396 rows x 20 columns]
```

[ ]:
```
# Diabetes,General_Health,Checkup,Age_Category mapping using ordinal encoding
diabetes={'No':0,'No, pre-diabetes or borderline diabetes':0,'Yes':1,'Yes, but␣
 ↪female told only during pregnancy':1}
health={'Poor':0,'Fair':1,'Good':2,'Very Good':3,'Excellent':4}
checkup={'Never':0,'5 or more years ago':0,'Within the past 5 years':1,'Within␣
 ↪the past 2 years':2,'Within the past year':3}
age={'18-24':0,'25-29':1,'30-34':2,'35-39':3,'40-44':4,'45-49':5,'50-54':
 ↪6,'55-59':7,'60-64':8,'65-69':9,'70-74':10,'75-79':11,'80+':12}
dfe['Diabetes']=dfe['Diabetes'].map(diabetes)
dfe['General_Health']=dfe['General_Health'].map(health)
dfe['Checkup']=dfe['Checkup'].map(checkup)
dfe['Age_Category']=dfe['Age_Category'].map(age)
dfe
```

[ ]:
```
        General_Health  Checkup  Exercise  Skin_Cancer  Other_Cancer  \
0                    0        2         0            0             0
1                    3        3         0            0             0
2                    3        3         1            0             0
3                    0        3         1            0             0
4                    2        3         0            0             0
…                  …      …       …          …           …
289391               3        3         1            0             0
289392               1        1         1            0             0
289393               3        0         1            0             0
289394               3        3         1            0             0
289395               4        3         1            0             0

        Depression  Diabetes  Arthritis  Age_Category  Height_(cm)  \
0                0         0          1            10          150
1                0         1          0            10          165
2                0         1          0             8          163
```

```
3                0          1           0            11           180
4                0          0           0            12           191
…              …        …         …           …            …
289391           0          0           0             1           168
289392           0          1           0             9           180
289393           1          1           0             2           157
289394           0          0           0             9           183
289395           0          0           0             5           160

        Weight_(kg)    BMI  Smoking_History  Alcohol_Consumption  \
0             32.66  14.54                1                    0
1             77.11  28.29                0                    0
2             88.45  33.47                0                    4
3             93.44  28.73                0                    0
4             88.45  24.37                1                    0
…             …    …               …                  …
289391        81.65  29.05                0                    4
289392        69.85  21.48                0                    8
289393        61.23  24.69                1                    4
289394        79.38  23.73                0                    3
289395        81.19  31.71                0                    1

        Fruit_Consumption  Green_Vegetables_Consumption  \
0                      30                            16
1                      30                             0
2                      12                             3
3                      30                            30
4                       8                             4
…                     …                           …
289391                 30                             8
289392                 15                            60
289393                 40                             8
289394                 30                            12
289395                  5                            12

        FriedPotato_Consumption  Heart_Disease  Sex_Female  Sex_Male
0                            12              0           1         0
1                             4              1           1         0
2                            16              0           1         0
3                             8              1           0         1
4                             0              0           0         1
…                           …            …         …       …
289391                        0              0           0         1
289392                        4              0           0         1
289393                        4              0           1         0
289394                        0              0           0         1
289395                        1              0           1         0
```

```
[289396 rows x 20 columns]
```

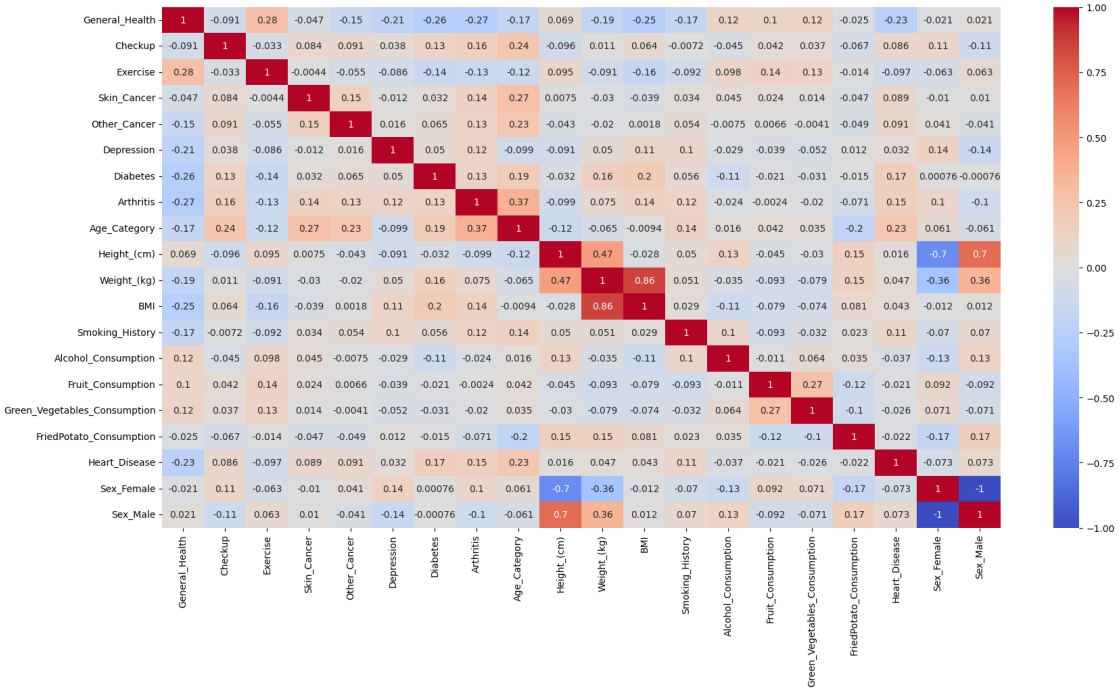
[ ]:
```
#Printing datatypes after encoding
dfe.dtypes
```

[ ]:
```
General_Health                  int64
Checkup                         int64
Exercise                        int64
Skin_Cancer                     int64
Other_Cancer                    int64
Depression                      int64
Diabetes                        int64
Arthritis                       int64
Age_Category                    int64
Height_(cm)                     int64
Weight_(kg)                   float64
BMI                           float64
Smoking_History                 int64
Alcohol_Consumption             int64
Fruit_Consumption               int64
Green_Vegetables_Consumption    int64
FriedPotato_Consumption         int64
Heart_Disease                   int64
Sex_Female                      uint8
Sex_Male                        uint8
dtype: object
```

CORRELATION MATRIX

[ ]:
```
plt.figure(figsize=(20,10))
sns.heatmap(dfe.corr(),annot=True,cmap='coolwarm')
```

[ ]:
```
<Axes: >
```



[ ]:
```
# To convert imbalaced dataset into balanced dataset
!pip install imbalanced-learn
```

```
Requirement already satisfied: imbalanced-learn in
/usr/local/lib/python3.10/dist-packages (0.10.1)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-
packages (from imbalanced-learn) (1.25.2)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-
packages (from imbalanced-learn) (1.11.4)
Requirement already satisfied: scikit-learn>=1.0.2 in
/usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (1.2.2)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-
packages (from imbalanced-learn) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (3.3.0)
```

[ ]:
```
import imblearn
```

SEPERATING X AND Y

[ ]:
```
x=dfe.drop('Heart_Disease',axis=1).values
x
```

[ ]:
```
array([[ 0.,  2.,  0., …, 12.,  1.,  0.],
       [ 3.,  3.,  0., …,  4.,  1.,  0.],
```

```
       [ 3.,  3.,  1., …, 16.,  1.,  0.],
       …,
       [ 3.,  0.,  1., …,  4.,  1.,  0.],
       [ 3.,  3.,  1., …,  0.,  0.,  1.],
       [ 4.,  3.,  1., …,  1.,  1.,  0.]])
```

[ ]:
```
y=dfe['Heart_Disease'].values
y
```

[ ]:
```
array([0, 1, 0, …, 0, 0, 0])
```

[ ]:
```
#Balncing the dataset
from imblearn.over_sampling import SMOTE
smote=SMOTE(sampling_strategy='minority',random_state=42)
x1,y1=smote.fit_resample(x,y)
```

[ ]:
```
#Coverting the dataset into training and testing data
from sklearn.model_selection import train_test_split
x1_train,x1_test,y1_train,y1_test=train_test_split(x1,y1,test_size=0.
 ↪30,random_state=42)
x1_train
```

[ ]:
```
array([[1.        , 3.        , 0.        , …, 9.48789024, 0.        ,
        1.        ],
       [2.40948112, 3.        , 1.        , …, 3.04740561, 0.        ,
        1.        ],
       [0.60596517, 3.        , 0.30298258, …, 8.        , 0.        ,
        1.        ],
       …,
       [2.62347286, 3.        , 1.        , …, 1.        , 1.        ,
        0.        ],
       [3.        , 3.        , 1.        , …, 4.        , 0.        ,
        1.        ],
       [4.        , 3.        , 1.        , …, 5.        , 0.        ,
        1.        ]])
```

[ ]:
```
x1_test
```

[ ]:
```
array([[2.        , 2.        , 1.        , …, 1.        , 0.        ,
        1.        ],
       [3.        , 2.        , 1.        , …, 3.        , 1.        ,
        0.        ],
       [2.        , 3.        , 1.        , …, 3.88651192, 0.        ,
        1.        ],
       …,
       [1.49323961, 3.        , 0.7466198 , …, 9.77366137, 0.2533802 ,
        0.7466198 ],
```

```
       [1.        , 3.        , 0.        , …, 0.        , 1.        ,
        0.        ],
       [4.        , 3.        , 1.        , …, 1.        , 0.        ,
        1.        ]])
```

[ ]:
```
y1_train
```

[ ]:
```
array([1, 1, 1, …, 1, 0, 0])
```

[ ]:
```
y1_test
```

[ ]:
```
array([0, 0, 1, …, 1, 0, 0])
```

NORMALIZATION

[ ]:
```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
scaler.fit(x1_train)
x1_train=scaler.transform(x1_train)
x1_test=scaler.transform(x1_test)
```

[ ]:
```
#Normalized training data
x1_train
```

[ ]:
```
array([[-1.11042222,  0.39598072, -1.73389825, …,  1.23858223,
        -0.97174703,  0.97174703],
       [ 0.22865983,  0.39598072,  0.6778443 , …, -0.34969063,
        -0.97174703,  0.97174703],
       [-1.48477628,  0.39598072, -1.00318226, …,  0.8716571 ,
        -0.97174703,  0.97174703],
       …,
       [ 0.43196336,  0.39598072,  0.6778443 , …, -0.85459654,
         1.09016136, -1.09016136],
       [ 0.78968418,  0.39598072,  0.6778443 , …, -0.11477355,
        -0.97174703,  0.97174703],
       [ 1.73973738,  0.39598072,  0.6778443 , …,  0.13183411,
        -0.97174703,  0.97174703]])
```

[ ]:
```
# Normalized testing data
x1_test
```

[ ]:
```
array([[-0.16036902, -1.1557933 ,  0.6778443 , …, -0.85459654,
        -0.97174703,  0.97174703],
       [ 0.78968418, -1.1557933 ,  0.6778443 , …, -0.36138121,
         1.09016136, -1.09016136],
       [-0.16036902,  0.39598072,  0.6778443 , …, -0.14276058,
        -0.97174703,  0.97174703],
       …,
```

```
       [-0.64181835,  0.39598072,  0.0667565 , …,  1.30905558,
        -0.44930028,  0.44930028],
       [-1.11042222,  0.39598072, -1.73389825, …, -1.1012042 ,
         1.09016136, -1.09016136],
       [ 1.73973738,  0.39598072,  0.6778443 , …, -0.85459654,
        -0.97174703,  0.97174703]])
```

MODEL CREATION

[ ]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import BernoulliNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import␣
 ↪accuracy_score,confusion_matrix,ConfusionMatrixDisplay,classification_report
knn=KNeighborsClassifier(n_neighbors=7)
bernoulli=BernoulliNB()
rfc=RandomForestClassifier(n_esimators=100,random_state=42)
```

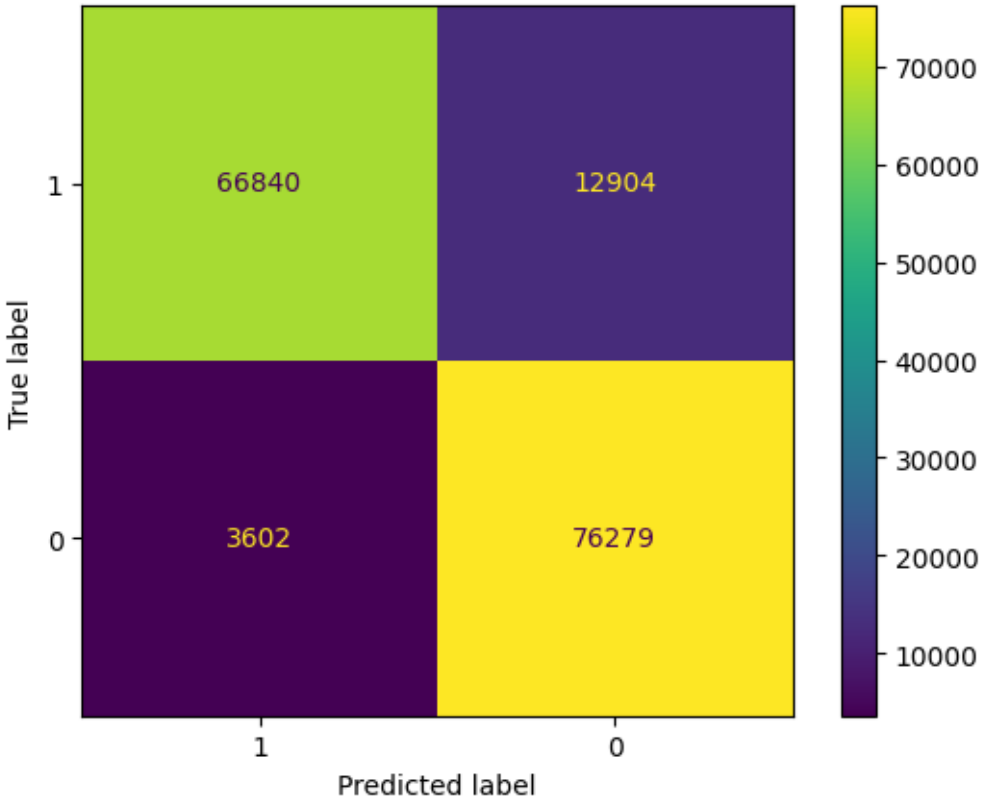
[ ]:
```
#KNearest Neighbors Algorithm
print('MODEL IS KNN')
knn.fit(x1_train,y1_train)
y1_pred1=knn.predict(x1_test)
print('SCORE IS:',accuracy_score(y1_test,y1_pred1))
print('-'*100)
cm=confusion_matrix(y1_test,y1_pred1)
print('MATRIX IS:',cm)
print('-'*100)
cmd=ConfusionMatrixDisplay(cm,display_labels=['1','0'])
print('MATRIX DISPLAY IS:',cmd.plot())
print('-'*100)
print('REPORT IS:',classification_report(y1_test,y1_pred1))
```

```
MODEL IS KNN
SCORE IS: 0.896595144870791
----------------------------------------------------------------------------------
--------------------
MATRIX IS: [[66840 12904]
 [ 3602 76279]]
----------------------------------------------------------------------------------
--------------------
MATRIX DISPLAY IS:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at
0x796f9d130700>
----------------------------------------------------------------------------------
--------------------
REPORT IS:               precision    recall  f1-score   support

           0       0.95      0.84      0.89     79744
```

```
              1       0.86      0.95      0.90     79881

       accuracy                           0.90    159625
      macro avg       0.90      0.90      0.90    159625
   weighted avg       0.90      0.90      0.90    159625
```



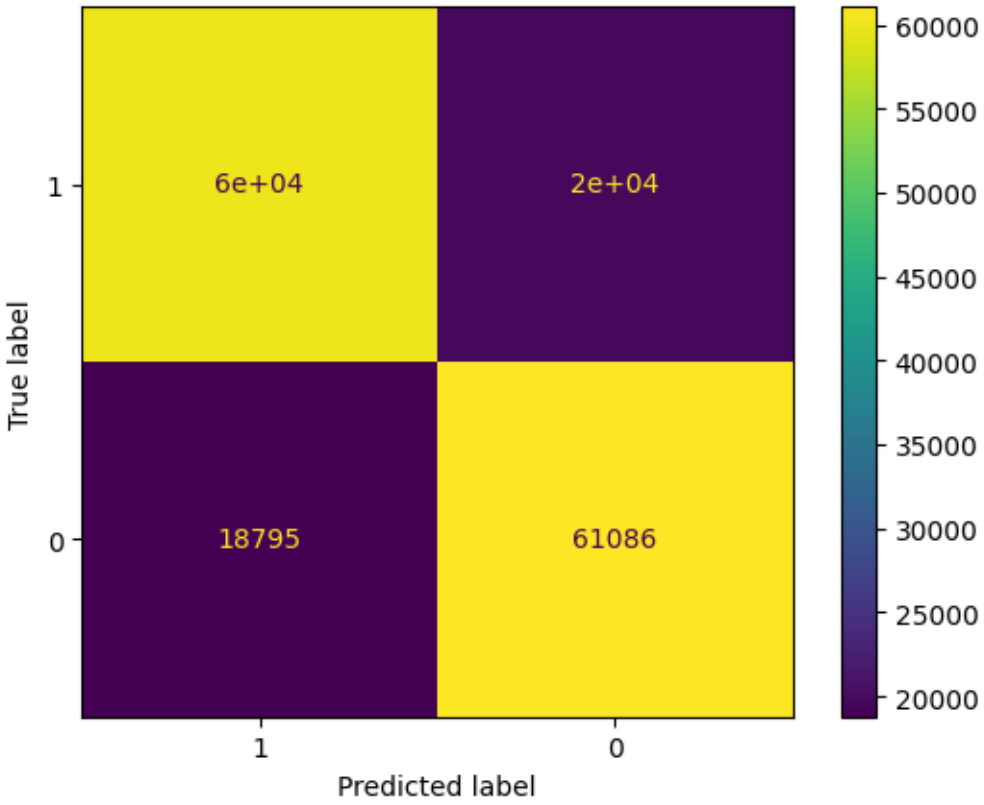
[ ]:
```
#Naive Bayes Algorithm
print('MODEL IS BernoulliNB')
bernoulli.fit(x1_train,y1_train)
y1_pred_bernoulli=bernoulli.predict(x1_test)
print('SCORE IS:',accuracy_score(y1_test,y1_pred_bernoulli))
print('-'*100)
cm_bernoulli=confusion_matrix(y1_test,y1_pred_bernoulli)
print('MATRIX IS:',cm_bernoulli)
print('-'*100)
cmd_bernoulli=ConfusionMatrixDisplay(cm_bernoulli,display_labels=['1','0'])
print('MATRIX DISPLAY IS:',cmd_bernoulli.plot())
print('-'*100)
print('REPORT IS:',classification_report(y1_test,y1_pred_bernoulli))
```

```
MODEL IS BernoulliNB
```

```
SCORE IS: 0.7584087705559907
--------------------------------------------------------------------------------
--------------------
MATRIX IS: [[59975 19769]
 [18795 61086]]
--------------------------------------------------------------------------------
--------------------
MATRIX DISPLAY IS:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at
0x796f9d1fff10>
--------------------------------------------------------------------------------
--------------------
REPORT IS:                precision    recall  f1-score   support

           0       0.76      0.75      0.76     79744
           1       0.76      0.76      0.76     79881

    accuracy                           0.76    159625
   macro avg       0.76      0.76      0.76    159625
weighted avg       0.76      0.76      0.76    159625
```

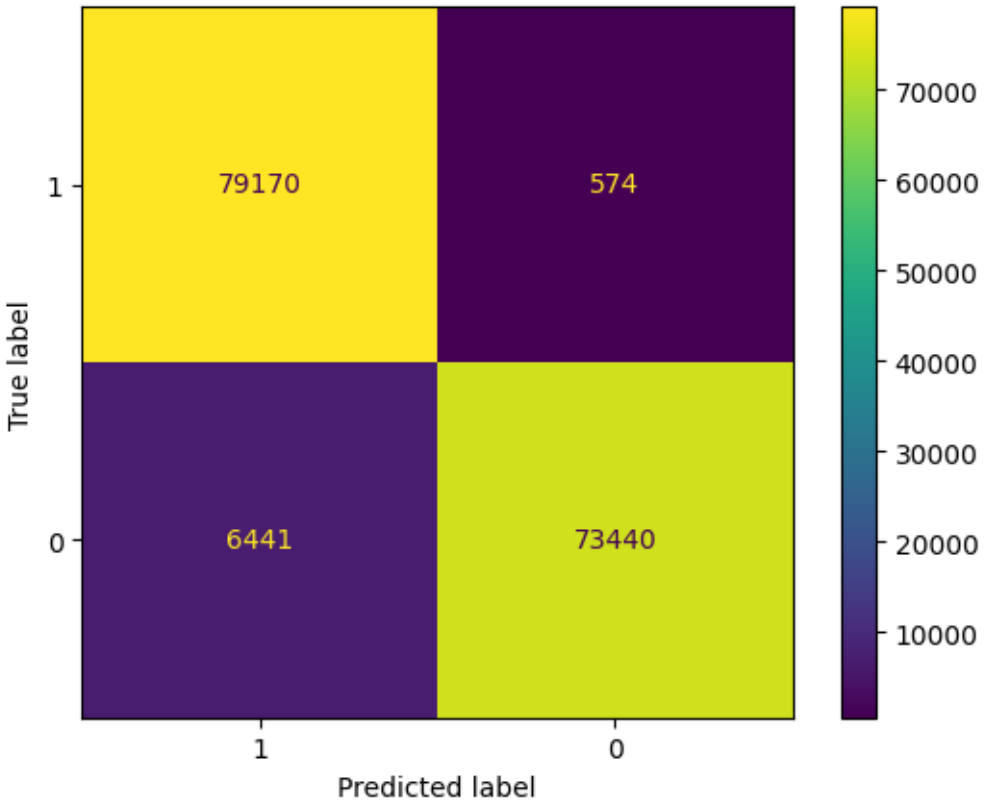


[ ]:
```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier(n_estimators=100,random_state=42)
rfc.fit(x1_train,y1_train)
rfc_pred=rfc.predict(x1_test)
print('SCORE IS:',accuracy_score(y1_test,rfc_pred))
print('-'*100)
cm_rfc=confusion_matrix(y1_test,rfc_pred)
print('MATRIX IS:',cm_rfc)
print('-'*100)
cmd_rfc=ConfusionMatrixDisplay(cm_rfc,display_labels=['1','0'])
print('MATRIX DISPLAY IS:',cmd_rfc.plot())
print('-'*100)
print('REPORT IS:',classification_report(y1_test,rfc_pred))
```

```
SCORE IS: 0.9560532498042287
----------------------------------------------------------------------------------------
--------------------
MATRIX IS: [[79170   574]
 [ 6441 73440]]
----------------------------------------------------------------------------------------
--------------------
MATRIX DISPLAY IS:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at
0x796f9d2c2500>
----------------------------------------------------------------------------------------
--------------------
REPORT IS:               precision    recall  f1-score   support

           0       0.92      0.99      0.96     79744
           1       0.99      0.92      0.95     79881

    accuracy                           0.96    159625
   macro avg       0.96      0.96      0.96    159625
weighted avg       0.96      0.96      0.96    159625
```



CONCLUSION

I successfully built a Random forest classifier model that can predict whether a person has heart disease with an accuracy of around 96%, Naive bayes model with 76% and KNeighbor classifier with 90% accuracy.

Here Random Forest Algorithm gives best performance.