05122022_Cervical_Cancer_Classification

December 6, 2022

0.1 Cervical Cancer Prediction Using XG-Boost

Cervical cancer is a type of cancer that affects the cervix, which is the lower part of the uterus that connects to the vagina. It is typically caused by the human papillomavirus (HPV), which is a sexually transmitted infection. Other contributing factors to the development of cervical cancer can include a weakened immune system, smoking, and having multiple sexual partners. Regular screening tests, such as Pap tests, can help detect cervical cancer early, when it is most treatable. Treatment options may include surgery, radiation therapy, and chemotherapy. It is important to reduce the impact of this disease because it can have serious health consequences. Early detection and treatment are key, as well as reducing the risk of developing cervical cancer by practicing safe sex and getting vaccinated against HPV.

0.1.1 Import necessary libraries and dataset.

[ ]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import zipfile
import plotly.express as px

from jupyterthemes import jtplot
jtplot.style(theme='monokai', context='notebook', ticks=True, grid=False)
```

[ ]:

```python
cancer_df = pd.read_csv('cervical_cancer.csv')
```

0.1.2 Understand the structure, format, data types of the dataframe by using .info() an .describe()

[ ]:

```python
cancer_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           858 non-null    object
```

```
  2   First sexual intercourse            858 non-null    object
  3   Num of pregnancies                  858 non-null    object
  4   Smokes                              858 non-null    object
  5   Smokes (years)                      858 non-null    object
  6   Smokes (packs/year)                 858 non-null    object
  7   Hormonal Contraceptives             858 non-null    object
  8   Hormonal Contraceptives (years)     858 non-null    object
  9   IUD                                 858 non-null    object
  10  IUD (years)                         858 non-null    object
  11  STDs                                858 non-null    object
  12  STDs (number)                       858 non-null    object
  13  STDs:condylomatosis                 858 non-null    object
  14  STDs:cervical condylomatosis        858 non-null    object
  15  STDs:vaginal condylomatosis         858 non-null    object
  16  STDs:vulvo-perineal condylomatosis  858 non-null    object
  17  STDs:syphilis                       858 non-null    object
  18  STDs:pelvic inflammatory disease    858 non-null    object
  19  STDs:genital herpes                 858 non-null    object
  20  STDs:molluscum contagiosum          858 non-null    object
  21  STDs:AIDS                           858 non-null    object
  22  STDs:HIV                            858 non-null    object
  23  STDs:Hepatitis B                    858 non-null    object
  24  STDs:HPV                            858 non-null    object
  25  STDs: Number of diagnosis           858 non-null    int64
  26  STDs: Time since first diagnosis    858 non-null    object
  27  STDs: Time since last diagnosis     858 non-null    object
  28  Dx:Cancer                           858 non-null    int64
  29  Dx:CIN                              858 non-null    int64
  30  Dx:HPV                              858 non-null    int64
  31  Dx                                  858 non-null    int64
  32  Hinselmann                          858 non-null    int64
  33  Schiller                            858 non-null    int64
  34  Citology                            858 non-null    int64
  35  Biopsy                              858 non-null    int64
dtypes: int64(10), object(26)
memory usage: 241.4+ KB
```

[ ]:

```
cancer_df.describe()
```

[ ]:

```
              Age  STDs: Number of diagnosis   Dx:Cancer      Dx:CIN  \
count  858.000000                 858.000000  858.000000  858.000000
mean    26.820513                   0.087413    0.020979    0.010490
std      8.497948                   0.302545    0.143398    0.101939
min     13.000000                   0.000000    0.000000    0.000000
25%     20.000000                   0.000000    0.000000    0.000000
50%     25.000000                   0.000000    0.000000    0.000000
75%     32.000000                   0.000000    0.000000    0.000000
```

```
max     84.000000                   3.000000    1.000000    1.000000

           Dx:HPV          Dx  Hinselmann    Schiller    Citology      Biopsy
count  858.000000  858.000000  858.000000  858.000000  858.000000  858.000000
mean     0.020979    0.027972    0.040793    0.086247    0.051282    0.064103
std      0.143398    0.164989    0.197925    0.280892    0.220701    0.245078
min      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
25%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
50%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
75%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
max      1.000000    1.000000    1.000000    1.000000    1.000000    1.000000
```

There are several vaues that are equal to ‘?’ we need to replace these so that we can start to analyse our data more clearly

[ ]:
```
cancer_df = cancer_df.replace('?', np.nan)
cancer_df
```

[ ]:
```
     Age Number of sexual partners First sexual intercourse  \
0     18                       4.0                     15.0
1     15                       1.0                     14.0
2     34                       1.0                      NaN
3     52                       5.0                     16.0
4     46                       3.0                     21.0
..   …                       …                      …
853   34                       3.0                     18.0
854   32                       2.0                     19.0
855   25                       2.0                     17.0
856   33                       2.0                     24.0
857   29                       2.0                     20.0

    Num of pregnancies Smokes Smokes (years) Smokes (packs/year)  \
0                  1.0    0.0            0.0                 0.0
1                  1.0    0.0            0.0                 0.0
2                  1.0    0.0            0.0                 0.0
3                  4.0    1.0           37.0                37.0
4                  4.0    0.0            0.0                 0.0
..                 …    …            …                 …
853                0.0    0.0            0.0                 0.0
854                1.0    0.0            0.0                 0.0
855                0.0    0.0            0.0                 0.0
856                2.0    0.0            0.0                 0.0
857                1.0    0.0            0.0                 0.0

    Hormonal Contraceptives Hormonal Contraceptives (years)  IUD  …  \
0                       0.0                             0.0  0.0  …
1                       0.0                             0.0  0.0  …
```

```
2                       0.0                             0.0  0.0  …
3                       1.0                             3.0  0.0  …
4                       1.0                            15.0  0.0  …
..                      …                             …  …  …
853                     0.0                             0.0  0.0  …
854                     1.0                             8.0  0.0  …
855                     1.0                            0.08  0.0  …
856                     1.0                            0.08  0.0  …
857                     1.0                             0.5  0.0  …

    STDs: Time since first diagnosis STDs: Time since last diagnosis  \
0                                NaN                             NaN
1                                NaN                             NaN
2                                NaN                             NaN
3                                NaN                             NaN
4                                NaN                             NaN
..                               …                             …
853                              NaN                             NaN
854                              NaN                             NaN
855                              NaN                             NaN
856                              NaN                             NaN
857                              NaN                             NaN

    Dx:Cancer Dx:CIN Dx:HPV Dx Hinselmann Schiller Citology Biopsy
0           0      0      0  0          0        0        0      0
1           0      0      0  0          0        0        0      0
2           0      0      0  0          0        0        0      0
3           1      0      1  0          0        0        0      0
4           0      0      0  0          0        0        0      0
..        …    …    … ..        …      …      …    …
853         0      0      0  0          0        0        0      0
854         0      0      0  0          0        0        0      0
855         0      0      0  0          0        0        1      0
856         0      0      0  0          0        0        0      0
857         0      0      0  0          0        0        0      0

[858 rows x 36 columns]
```

[ ]:

```
cancer_df.isnull()
```

[ ]:

```
       Age  Number of sexual partners  First sexual intercourse  \
0    False                      False                     False
1    False                      False                     False
2    False                      False                      True
3    False                      False                     False
4    False                      False                     False
..     …                        …                       …
```

```
853  False                      False                     False
854  False                      False                     False
855  False                      False                     False
856  False                      False                     False
857  False                      False                     False

     Num of pregnancies  Smokes  Smokes (years)  Smokes (packs/year)  \
0                 False   False           False                False
1                 False   False           False                False
2                 False   False           False                False
3                 False   False           False                False
4                 False   False           False                False
..                  …     …             …                  …
853               False   False           False                False
854               False   False           False                False
855               False   False           False                False
856               False   False           False                False
857               False   False           False                False

     Hormonal Contraceptives  Hormonal Contraceptives (years)    IUD  …  \
0                      False                            False  False  …
1                      False                            False  False  …
2                      False                            False  False  …
3                      False                            False  False  …
4                      False                            False  False  …
..                       …                              …    …  …
853                    False                            False  False  …
854                    False                            False  False  …
855                    False                            False  False  …
856                    False                            False  False  …
857                    False                            False  False  …

     STDs: Time since first diagnosis  STDs: Time since last diagnosis  \
0                                True                             True
1                                True                             True
2                                True                             True
3                                True                             True
4                                True                             True
..                                …                              …
853                              True                             True
854                              True                             True
855                              True                             True
856                              True                             True
857                              True                             True

     Dx:Cancer  Dx:CIN  Dx:HPV     Dx  Hinselmann  Schiller  Citology  Biopsy
0        False   False   False  False       False     False     False   False
```

```
1        False   False   False  False       False     False     False   False
2        False   False   False  False       False     False     False   False
3        False   False   False  False       False     False     False   False
4        False   False   False  False       False     False     False   False
..          …    …     …    …         …       …       …     …
853      False   False   False  False       False     False     False   False
854      False   False   False  False       False     False     False   False
855      False   False   False  False       False     False     False   False
856      False   False   False  False       False     False     False   False
857      False   False   False  False       False     False     False   False

[858 rows x 36 columns]
```

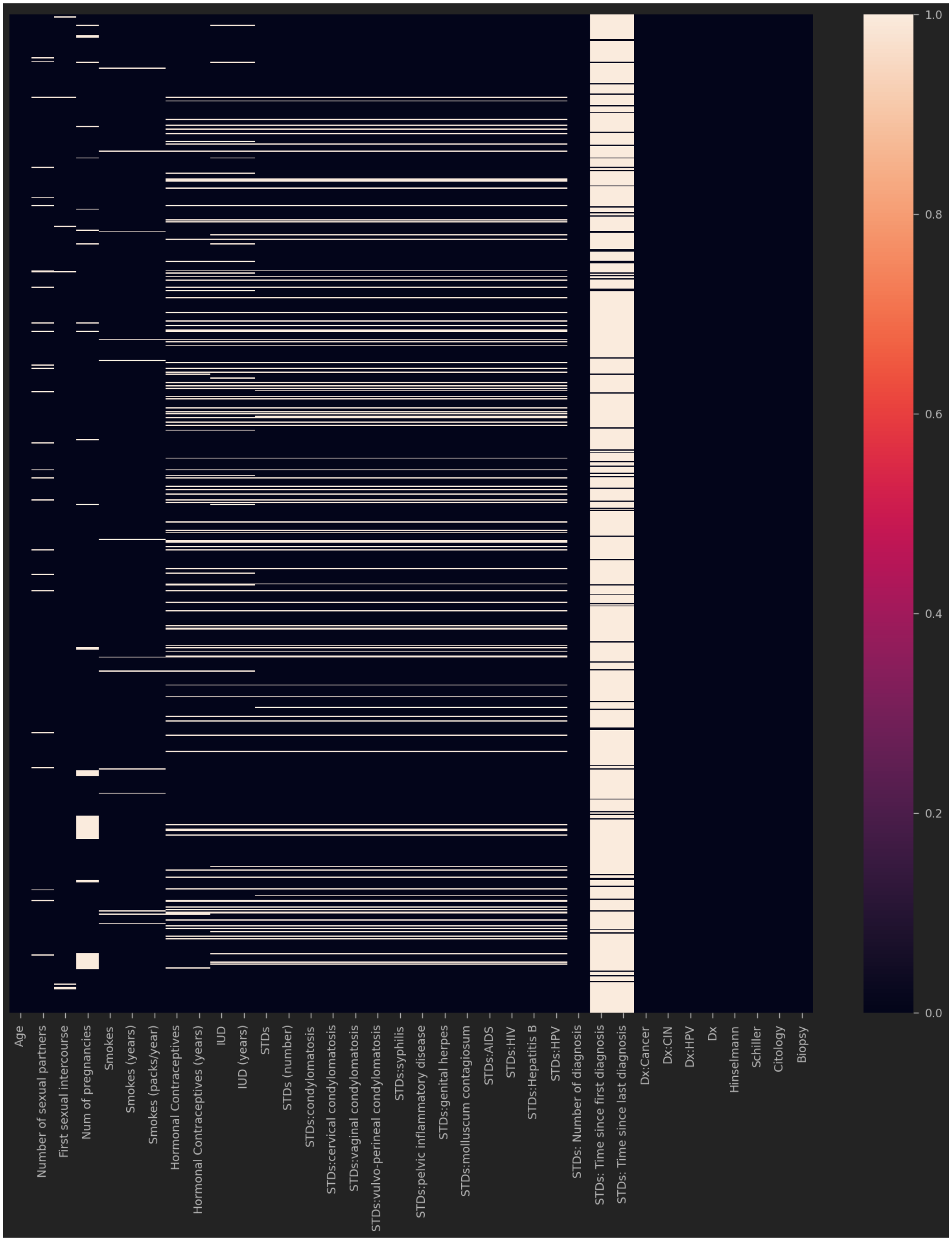
Create a heatmap to represent how many null values are present in the dataset.

[ ]:

```
plt.figure(figsize= (20, 20))
sns.heatmap(cancer_df.isnull(), yticklabels=False)
```

[ ]:

```
<AxesSubplot:>
```



It would be wise to drop the two columns with high amounts of null since we do not actually have a lot of data we can work with when we discount the null values.

[ ]:

```
cancer_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           832 non-null    object
 2   First sexual intercourse            851 non-null    object
 3   Num of pregnancies                  802 non-null    object
 4   Smokes                              845 non-null    object
 5   Smokes (years)                      845 non-null    object
 6   Smokes (packs/year)                 845 non-null    object
 7   Hormonal Contraceptives             750 non-null    object
 8   Hormonal Contraceptives (years)     750 non-null    object
 9   IUD                                 741 non-null    object
 10  IUD (years)                         741 non-null    object
 11  STDs                                753 non-null    object
 12  STDs (number)                       753 non-null    object
 13  STDs:condylomatosis                 753 non-null    object
 14  STDs:cervical condylomatosis        753 non-null    object
 15  STDs:vaginal condylomatosis         753 non-null    object
 16  STDs:vulvo-perineal condylomatosis  753 non-null    object
 17  STDs:syphilis                       753 non-null    object
 18  STDs:pelvic inflammatory disease    753 non-null    object
 19  STDs:genital herpes                 753 non-null    object
 20  STDs:molluscum contagiosum          753 non-null    object
 21  STDs:AIDS                           753 non-null    object
 22  STDs:HIV                            753 non-null    object
 23  STDs:Hepatitis B                    753 non-null    object
 24  STDs:HPV                            753 non-null    object
 25  STDs: Number of diagnosis           858 non-null    int64
 26  STDs: Time since first diagnosis    71 non-null     object
 27  STDs: Time since last diagnosis     71 non-null     object
 28  Dx:Cancer                           858 non-null    int64
 29  Dx:CIN                              858 non-null    int64
 30  Dx:HPV                              858 non-null    int64
 31  Dx                                  858 non-null    int64
 32  Hinselmann                          858 non-null    int64
 33  Schiller                            858 non-null    int64
 34  Citology                            858 non-null    int64
 35  Biopsy                              858 non-null    int64
dtypes: int64(10), object(26)
memory usage: 241.4+ KB
```

[ ]:

```
cancer_df = cancer_df.drop(columns= ['STDs: Time since first diagnosis','STDs:␣
↪Time since last diagnosis'])
cancer_df
```

```
[ ]:      Age Number of sexual partners First sexual intercourse  \
     0     18                        4.0                      15.0
     1     15                        1.0                      14.0
     2     34                        1.0                       NaN
     3     52                        5.0                      16.0
     4     46                        3.0                      21.0
     ..   …                        …                      …
     853   34                        3.0                      18.0
     854   32                        2.0                      19.0
     855   25                        2.0                      17.0
     856   33                        2.0                      24.0
     857   29                        2.0                      20.0

         Num of pregnancies Smokes Smokes (years) Smokes (packs/year)  \
     0                  1.0    0.0            0.0                 0.0
     1                  1.0    0.0            0.0                 0.0
     2                  1.0    0.0            0.0                 0.0
     3                  4.0    1.0           37.0                37.0
     4                  4.0    0.0            0.0                 0.0
     ..                  …    …            …                 …
     853                0.0    0.0            0.0                 0.0
     854                1.0    0.0            0.0                 0.0
     855                0.0    0.0            0.0                 0.0
     856                2.0    0.0            0.0                 0.0
     857                1.0    0.0            0.0                 0.0

         Hormonal Contraceptives Hormonal Contraceptives (years)  IUD  …  \
     0                       0.0                             0.0  0.0  …
     1                       0.0                             0.0  0.0  …
     2                       0.0                             0.0  0.0  …
     3                       1.0                             3.0  0.0  …
     4                       1.0                            15.0  0.0  …
     ..                      …                            …  …  …
     853                     0.0                             0.0  0.0  …
     854                     1.0                             8.0  0.0  …
     855                     1.0                            0.08  0.0  …
     856                     1.0                            0.08  0.0  …
     857                     1.0                             0.5  0.0  …

         STDs:HPV STDs: Number of diagnosis Dx:Cancer Dx:CIN Dx:HPV Dx Hinselmann  \
     0        0.0                         0         0      0      0  0          0
     1        0.0                         0         0      0      0  0          0
     2        0.0                         0         0      0      0  0          0
     3        0.0                         0         1      0      1  0          0
     4        0.0                         0         0      0      0  0          0
     ..       …                        …       …    …    … ..        …
     853      0.0                         0         0      0      0  0          0
```

```
854      0.0                          0         0      0      0  0          0
855      0.0                          0         0      0      0  0          0
856      0.0                          0         0      0      0  0          0
857      0.0                          0         0      0      0  0          0

    Schiller Citology Biopsy
0          0        0      0
1          0        0      0
2          0        0      0
3          0        0      0
4          0        0      0
..       …      …    …
853        0        0      0
854        0        0      0
855        0        1      0
856        0        0      0
857        0        0      0

[858 rows x 34 columns]
```

We should also convert our object data types to a numeric type.

[ ]:
```
cancer_df = cancer_df.apply(pd.to_numeric)
cancer_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 34 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           832 non-null    float64
 2   First sexual intercourse            851 non-null    float64
 3   Num of pregnancies                  802 non-null    float64
 4   Smokes                              845 non-null    float64
 5   Smokes (years)                      845 non-null    float64
 6   Smokes (packs/year)                 845 non-null    float64
 7   Hormonal Contraceptives             750 non-null    float64
 8   Hormonal Contraceptives (years)     750 non-null    float64
 9   IUD                                 741 non-null    float64
 10  IUD (years)                         741 non-null    float64
 11  STDs                                753 non-null    float64
 12  STDs (number)                       753 non-null    float64
 13  STDs:condylomatosis                 753 non-null    float64
 14  STDs:cervical condylomatosis        753 non-null    float64
 15  STDs:vaginal condylomatosis         753 non-null    float64
 16  STDs:vulvo-perineal condylomatosis  753 non-null    float64
 17  STDs:syphilis                       753 non-null    float64
 18  STDs:pelvic inflammatory disease    753 non-null    float64
 19  STDs:genital herpes                 753 non-null    float64
 20  STDs:molluscum contagiosum          753 non-null    float64
 21  STDs:AIDS                           753 non-null    float64
 22  STDs:HIV                            753 non-null    float64
 23  STDs:Hepatitis B                    753 non-null    float64
 24  STDs:HPV                            753 non-null    float64
 25  STDs: Number of diagnosis           858 non-null    int64
 26  Dx:Cancer                           858 non-null    int64
 27  Dx:CIN                              858 non-null    int64
 28  Dx:HPV                              858 non-null    int64
 29  Dx                                  858 non-null    int64
 30  Hinselmann                          858 non-null    int64
 31  Schiller                            858 non-null    int64
 32  Citology                            858 non-null    int64
 33  Biopsy                              858 non-null    int64
dtypes: float64(24), int64(10)
memory usage: 228.0 KB
```

[ ]:

```
cancer_df.describe()
```

[ ]:

```
              Age  Number of sexual partners  First sexual intercourse  \
count  858.000000                 832.000000                851.000000   
mean    26.820513                   2.527644                 16.995300   
std      8.497948                   1.667760                  2.803355   
min     13.000000                   1.000000                 10.000000   
25%     20.000000                   2.000000                 15.000000   
50%     25.000000                   2.000000                 17.000000   
75%     32.000000                   3.000000                 18.000000   
max     84.000000                  28.000000                 32.000000   

       Num of pregnancies      Smokes  Smokes (years)  Smokes (packs/year)  \
count          802.000000  845.000000      845.000000           845.000000   
mean             2.275561    0.145562        1.219721             0.453144   
std              1.447414    0.352876        4.089017             2.226610   
min              0.000000    0.000000        0.000000             0.000000   
25%              1.000000    0.000000        0.000000             0.000000   
50%              2.000000    0.000000        0.000000             0.000000   
75%              3.000000    0.000000        0.000000             0.000000   
max             11.000000    1.000000       37.000000            37.000000   

       Hormonal Contraceptives  Hormonal Contraceptives (years)         IUD  \
count               750.000000                       750.000000  741.000000   
mean                  0.641333                         2.256419    0.112011   
std                   0.479929                         3.764254    0.315593   
min                   0.000000                         0.000000    0.000000   
25%                   0.000000                         0.000000    0.000000   
```

```
50%                    1.000000                             0.500000    0.000000
75%                    1.000000                             3.000000    0.000000
max                    1.000000                            30.000000    1.000000

       …   STDs:HPV  STDs: Number of diagnosis   Dx:Cancer      Dx:CIN  \
count  …  753.000000                858.000000  858.000000  858.000000
mean   …    0.002656                  0.087413    0.020979    0.010490
std    …    0.051503                  0.302545    0.143398    0.101939
min    …    0.000000                  0.000000    0.000000    0.000000
25%    …    0.000000                  0.000000    0.000000    0.000000
50%    …    0.000000                  0.000000    0.000000    0.000000
75%    …    0.000000                  0.000000    0.000000    0.000000
max    …    1.000000                  3.000000    1.000000    1.000000

           Dx:HPV          Dx  Hinselmann    Schiller    Citology      Biopsy
count  858.000000  858.000000  858.000000  858.000000  858.000000  858.000000
mean     0.020979    0.027972    0.040793    0.086247    0.051282    0.064103
std      0.143398    0.164989    0.197925    0.280892    0.220701    0.245078
min      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
25%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
50%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
75%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
max      1.000000    1.000000    1.000000    1.000000    1.000000    1.000000

[8 rows x 34 columns]
```

We will now replace our null values in other columns with the average for that column. We can then use a heatmap to discover if we have any remaining null values.

[ ]:
```
cancer_df.mean()
```

[ ]:
```
Age                                   26.820513
Number of sexual partners              2.527644
First sexual intercourse              16.995300
Num of pregnancies                     2.275561
Smokes                                 0.145562
Smokes (years)                         1.219721
Smokes (packs/year)                    0.453144
Hormonal Contraceptives                0.641333
Hormonal Contraceptives (years)        2.256419
IUD                                    0.112011
IUD (years)                            0.514804
STDs                                   0.104914
STDs (number)                          0.176627
STDs:condylomatosis                    0.058433
STDs:cervical condylomatosis           0.000000
STDs:vaginal condylomatosis            0.005312
```

```
STDs:vulvo-perineal condylomatosis     0.057105
STDs:syphilis                          0.023904
STDs:pelvic inflammatory disease       0.001328
STDs:genital herpes                    0.001328
STDs:molluscum contagiosum             0.001328
STDs:AIDS                              0.000000
STDs:HIV                               0.023904
STDs:Hepatitis B                       0.001328
STDs:HPV                               0.002656
STDs: Number of diagnosis              0.087413
Dx:Cancer                              0.020979
Dx:CIN                                 0.010490
Dx:HPV                                 0.020979
Dx                                     0.027972
Hinselmann                             0.040793
Schiller                               0.086247
Citology                               0.051282
Biopsy                                 0.064103
dtype: float64
```

[ ]:
```
cancer_df = cancer_df.fillna(cancer_df.mean())
cancer_df
```

[ ]:
```
     Age  Number of sexual partners  First sexual intercourse  \
0     18                        4.0                   15.0000
1     15                        1.0                   14.0000
2     34                        1.0                   16.9953
3     52                        5.0                   16.0000
4     46                        3.0                   21.0000
..   …                        …                       …
853   34                        3.0                   18.0000
854   32                        2.0                   19.0000
855   25                        2.0                   17.0000
856   33                        2.0                   24.0000
857   29                        2.0                   20.0000

     Num of pregnancies  Smokes  Smokes (years)  Smokes (packs/year)  \
0                   1.0     0.0             0.0                  0.0
1                   1.0     0.0             0.0                  0.0
2                   1.0     0.0             0.0                  0.0
3                   4.0     1.0            37.0                 37.0
4                   4.0     0.0             0.0                  0.0
..                  …     …             …                  …
853                 0.0     0.0             0.0                  0.0
854                 1.0     0.0             0.0                  0.0
855                 0.0     0.0             0.0                  0.0
856                 2.0     0.0             0.0                  0.0
```

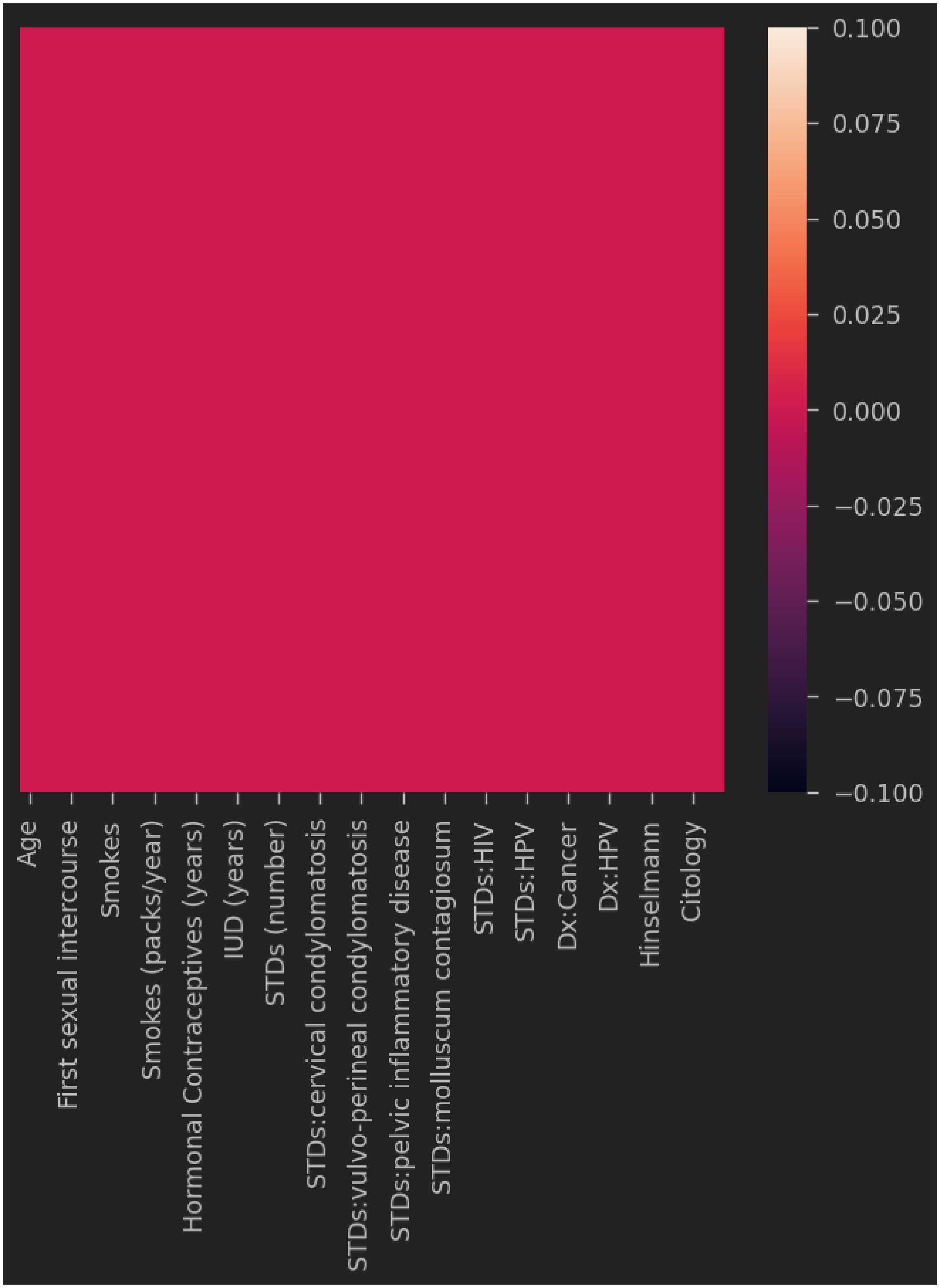
```
857                 1.0     0.0             0.0                  0.0

     Hormonal Contraceptives  Hormonal Contraceptives (years)  IUD  …  \
0                        0.0                             0.00  0.0  …
1                        0.0                             0.00  0.0  …
2                        0.0                             0.00  0.0  …
3                        1.0                             3.00  0.0  …
4                        1.0                            15.00  0.0  …
..                       …                              …  …  …
853                      0.0                             0.00  0.0  …
854                      1.0                             8.00  0.0  …
855                      1.0                             0.08  0.0  …
856                      1.0                             0.08  0.0  …
857                      1.0                             0.50  0.0  …

     STDs:HPV  STDs: Number of diagnosis  Dx:Cancer  Dx:CIN  Dx:HPV  Dx  \
0         0.0                          0          0       0       0   0
1         0.0                          0          0       0       0   0
2         0.0                          0          0       0       0   0
3         0.0                          0          1       0       1   0
4         0.0                          0          0       0       0   0
..        …                        …        …     …     …  ..
853       0.0                          0          0       0       0   0
854       0.0                          0          0       0       0   0
855       0.0                          0          0       0       0   0
856       0.0                          0          0       0       0   0
857       0.0                          0          0       0       0   0

     Hinselmann  Schiller  Citology  Biopsy
0             0         0         0       0
1             0         0         0       0
2             0         0         0       0
3             0         0         0       0
4             0         0         0       0
..          …       …       …     …
853           0         0         0       0
854           0         0         0       0
855           0         0         1       0
856           0         0         0       0
857           0         0         0       0

[858 rows x 34 columns]
```

[ ]:

```python
sns.heatmap(cancer_df.isnull(), yticklabels=False)
```

[ ]:

```
<AxesSubplot:>
```



Since we have one color, this indicates that there are no null values remaining.

[ ]:

```
min = cancer_df['Age'].min()
max = cancer_df['Age'].max()


def age_range(min, max):
    age_range = max - min
    return age_range


print(age_range(min, max))
```

```
71
```

[ ]:

```
cancer_df[cancer_df['Age'] == 84]
```

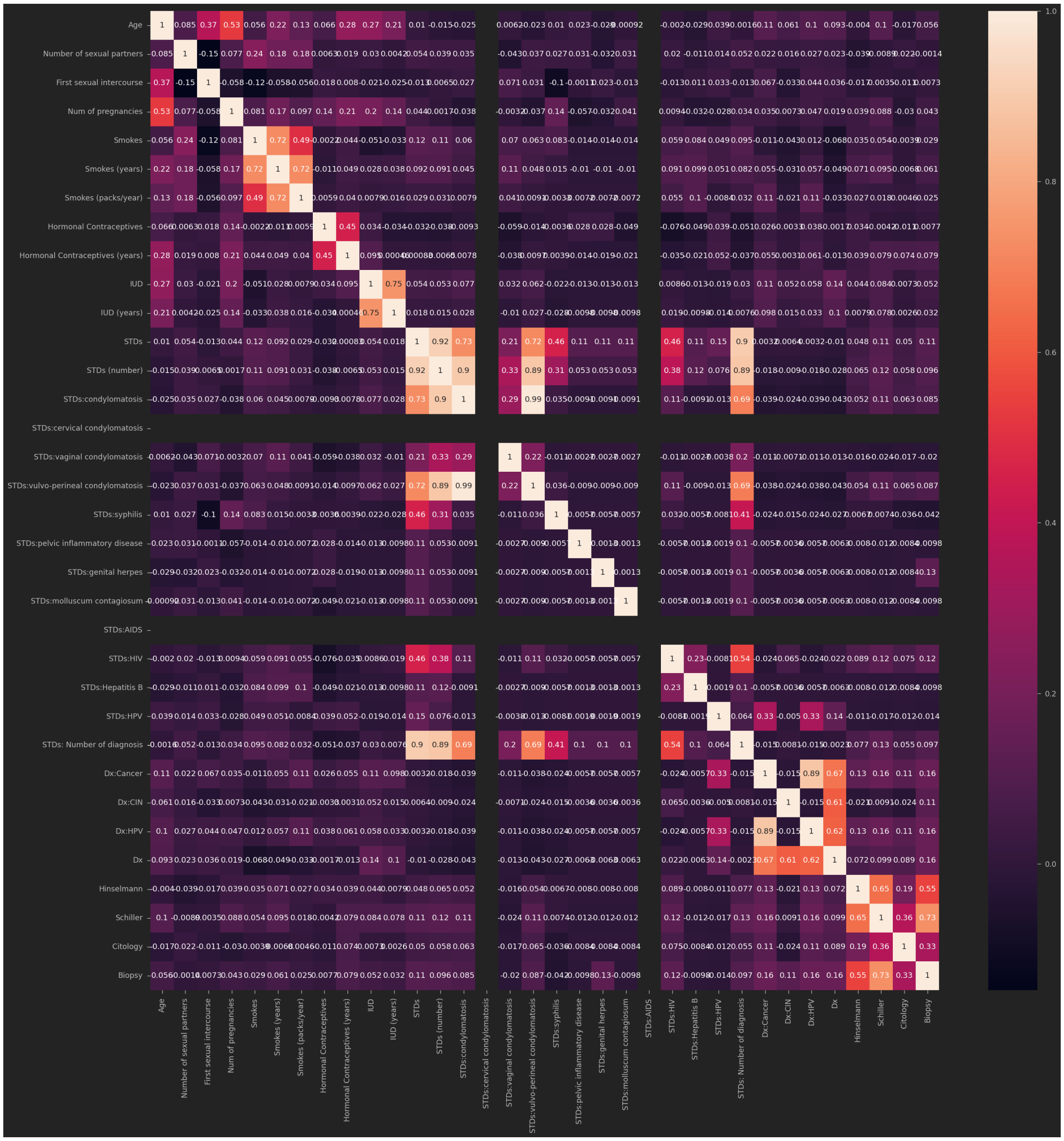
[ ]:

```
     Age  Number of sexual partners  First sexual intercourse  \
668   84                        3.0                      20.0

     Num of pregnancies  Smokes  Smokes (years)  Smokes (packs/year)  \
668                11.0     1.0            24.0             0.513202

     Hormonal Contraceptives  Hormonal Contraceptives (years)  IUD  …  \
668                      0.0                              0.0  0.0  …

     STDs:HPV  STDs: Number of diagnosis  Dx:Cancer  Dx:CIN  Dx:HPV  Dx  \
668       0.0                          0          0       0       0   0

     Hinselmann  Schiller  Citology  Biopsy
668           0         0         0       0

[1 rows x 34 columns]
```

0.1.3 Data Visualisation

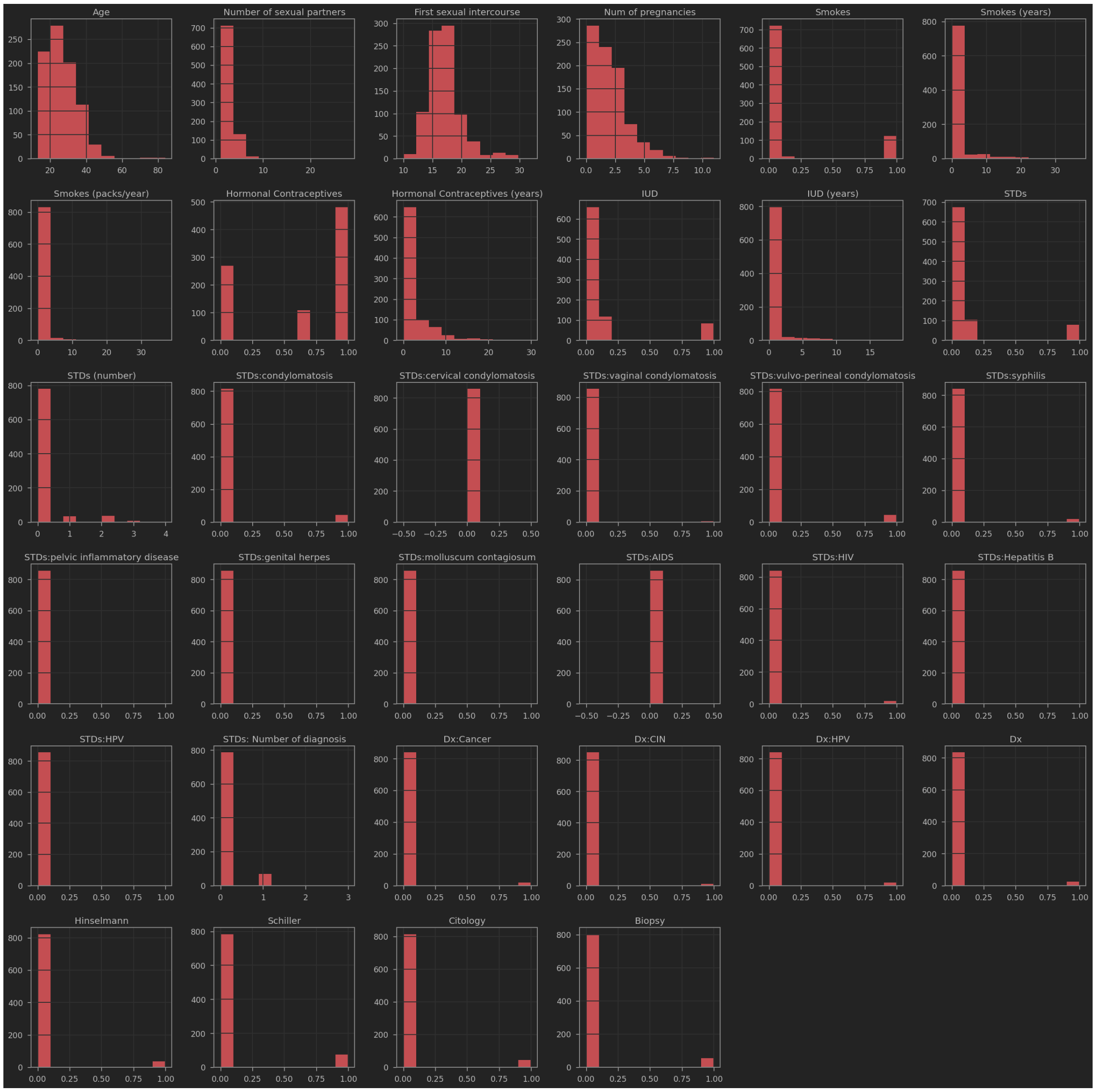
[ ]:

```
corr = cancer_df.corr()

plt.figure(figsize = (30,30))
sns.heatmap(corr, annot=True)
plt.show()
```



```
[ ]: cancer_df.hist(bins=10, figsize=(30, 30), color='r')
```

```
[ ]: array([[<AxesSubplot:title={'center':'Age'}>,
             <AxesSubplot:title={'center':'Number of sexual partners'}>,
             <AxesSubplot:title={'center':'First sexual intercourse'}>,
             <AxesSubplot:title={'center':'Num of pregnancies'}>,
             <AxesSubplot:title={'center':'Smokes'}>,
             <AxesSubplot:title={'center':'Smokes (years)'}>],
            [<AxesSubplot:title={'center':'Smokes (packs/year)'}>,
             <AxesSubplot:title={'center':'Hormonal Contraceptives'}>,
             <AxesSubplot:title={'center':'Hormonal Contraceptives (years)'}>,
             <AxesSubplot:title={'center':'IUD'}>,
             <AxesSubplot:title={'center':'IUD (years)'}>,
             <AxesSubplot:title={'center':'STDs'}>],
            [<AxesSubplot:title={'center':'STDs (number)'}>,
             <AxesSubplot:title={'center':'STDs:condylomatosis'}>,
             <AxesSubplot:title={'center':'STDs:cervical condylomatosis'}>,
             <AxesSubplot:title={'center':'STDs:vaginal condylomatosis'}>,
             <AxesSubplot:title={'center':'STDs:vulvo-perineal condylomatosis'}>,
             <AxesSubplot:title={'center':'STDs:syphilis'}>],
            [<AxesSubplot:title={'center':'STDs:pelvic inflammatory disease'}>,
             <AxesSubplot:title={'center':'STDs:genital herpes'}>,
             <AxesSubplot:title={'center':'STDs:molluscum contagiosum'}>,
             <AxesSubplot:title={'center':'STDs:AIDS'}>,
             <AxesSubplot:title={'center':'STDs:HIV'}>,
             <AxesSubplot:title={'center':'STDs:Hepatitis B'}>],
            [<AxesSubplot:title={'center':'STDs:HPV'}>,
             <AxesSubplot:title={'center':'STDs: Number of diagnosis'}>,
             <AxesSubplot:title={'center':'Dx:Cancer'}>,
             <AxesSubplot:title={'center':'Dx:CIN'}>,
             <AxesSubplot:title={'center':'Dx:HPV'}>,
             <AxesSubplot:title={'center':'Dx'}>],
            [<AxesSubplot:title={'center':'Hinselmann'}>,
             <AxesSubplot:title={'center':'Schiller'}>,
             <AxesSubplot:title={'center':'Citology'}>,
             <AxesSubplot:title={'center':'Biopsy'}>, <AxesSubplot:>,
             <AxesSubplot:>]], dtype=object)
```



Since we have visualised our data, we will now prepare our data to be used in model training. This is where things get more fun.

[ ]:
```
target_df = cancer_df['Biopsy']
input_df = cancer_df.drop(columns=['Biopsy'])
```

[ ]:
```
X = np.array(input_df).astype('float32')
y = np.array(target_df).astype('float32')
```

[ ]:
```
#we need to reshape the array but this version of python does it for us.
y.shape
```

[ ]:
```
(858,)
```

[ ]:

```
#applies a form of regularisation to our dataset.
from sklearn.preprocessing import StandardScaler, MinMaxScaler
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

[ ]:

```
#splitting the data in to test and train sets
from sklearn.model_selection import train_test_split

#assinging 40% of the data to testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.4)
#assigning 50% of the 40% of data assigned to test to be assigned to validation
X_test, X_val, y_test, y_val = train_test_split(X_test, y_test, test_size=.5)
```

[ ]:

```
#checking to see if I already have xgboost
# !pip install xgboost
```

[ ]:

```
#import and training XGBoost model.
import xgboost as xgb

model = xgb.XGBClassifier(learning_rate=.1, max_depth=25, n_estimators=100)

model.fit(X_train, y_train)
```

[ ]:

```
XGBClassifier(base_score=0.5, booster='gbtree', callbacks=None,
              colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
              early_stopping_rounds=None, enable_categorical=False,
              eval_metric=None, feature_types=None, gamma=0, gpu_id=-1,
              grow_policy='depthwise', importance_type=None,
              interaction_constraints='', learning_rate=0.1, max_bin=256,
              max_cat_threshold=64, max_cat_to_onehot=4, max_delta_step=0,
              max_depth=25, max_leaves=0, min_child_weight=1, missing=nan,
              monotone_constraints='()', n_estimators=100, n_jobs=0,
              num_parallel_tree=1, predictor='auto', random_state=0, …)
```

[ ]:

```
result_train = model.score(X_train, y_train)
result_train
```

[ ]:

```
0.9980544747081712
```

[ ]:

```
result_test = model.score(X_test, y_test)
result_test
```

[ ]:

```
0.9476744186046512
```

[ ]:

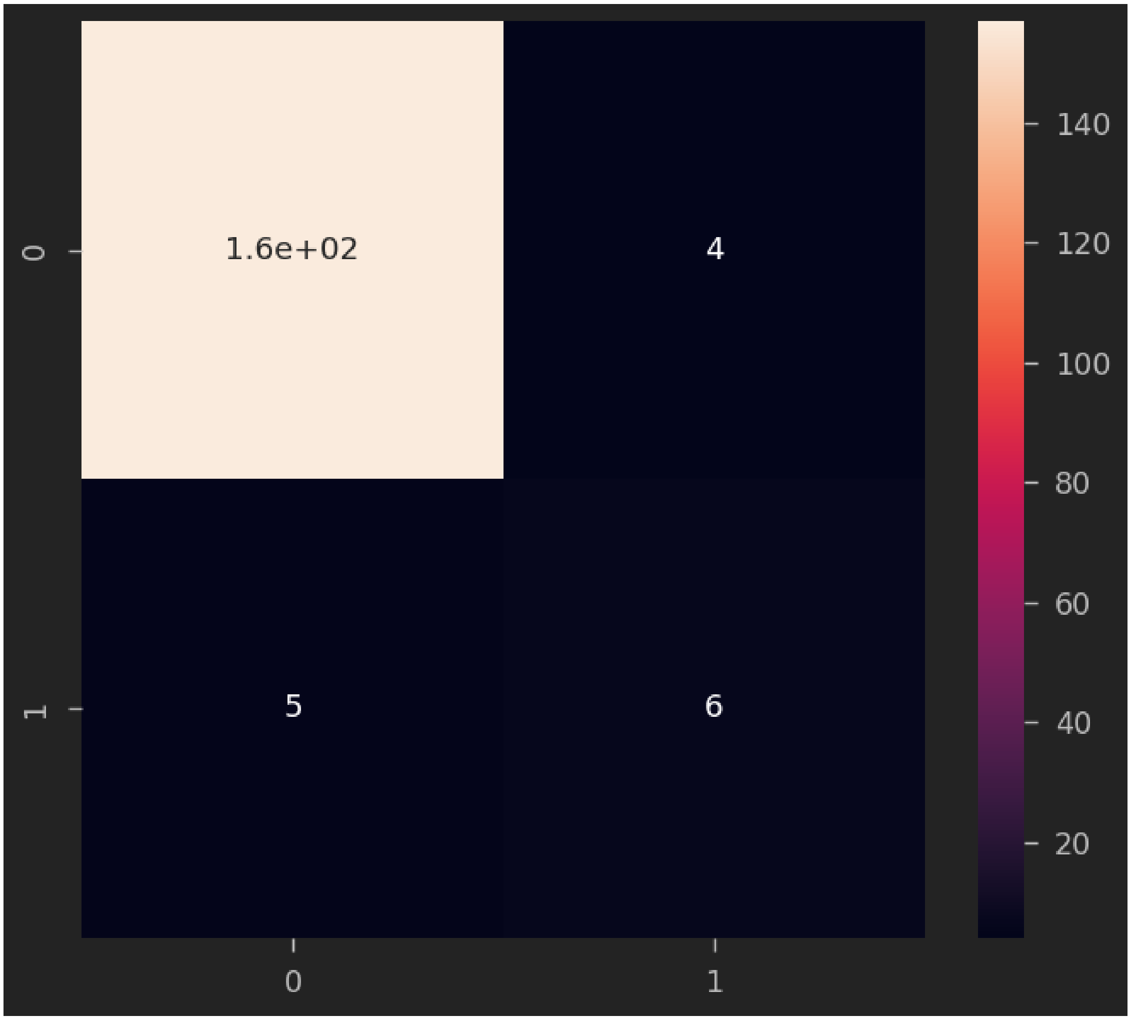
```
#make predictions on the test data
y_predict = model.predict(X_test)
```

```
[ ]: from sklearn.metrics import confusion_matrix, classification_report
     print(classification_report(y_test, y_predict))
```

```
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.97       162
         1.0       0.55      0.60      0.57        10

    accuracy                           0.95       172
   macro avg       0.76      0.78      0.77       172
weighted avg       0.95      0.95      0.95       172
```

```
[ ]: cm = confusion_matrix(y_predict, y_test)
     sns.heatmap(cm, annot=True)
```

```
[ ]: <AxesSubplot:>
```


1.6e+02
4
5
6
0
1
0
1
140
120
100
80
60
40
20