cervical-cancer-prediction

October 24, 2024

Cervical_Cancer_Prediction Project **Soham Katlariwala** are available @ **GitHub**

1 Predicting Cervical Cancer

1.0.1 STEP 1: IMPORTING LIBRARIES AND DATASETS

[ ]:

```
!pip install pandas
!pip install numpy
!pip install seaborn

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import zipfile

!pip install plotly
!pip install jupyterthemes
import plotly.express as px

from jupyterthemes import jtplot

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

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages
(2.2.2)
Requirement already satisfied: numpy>=1.22.4 in /usr/local/lib/python3.10/dist-
packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-
packages (from pandas) (2024.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

```
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages
(1.26.4)
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-
packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in
/usr/local/lib/python3.10/dist-packages (from seaborn) (1.26.4)
Requirement already satisfied: pandas>=1.2 in /usr/local/lib/python3.10/dist-
packages (from seaborn) (2.2.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in
/usr/local/lib/python3.10/dist-packages (from seaborn) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(1.3.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-
packages (from matplotlib!=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(4.54.1)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(1.4.7)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(24.1)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-
packages (from matplotlib!=3.6.1,>=3.4->seaborn) (10.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(3.1.4)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.10/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas>=1.2->seaborn) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-
packages (from pandas>=1.2->seaborn) (2024.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn) (1.16.0)
Requirement already satisfied: plotly in /usr/local/lib/python3.10/dist-packages
(5.24.1)
Requirement already satisfied: tenacity>=6.2.0 in
/usr/local/lib/python3.10/dist-packages (from plotly) (9.0.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from plotly) (24.1)
Collecting jupyterthemes
  Downloading jupyterthemes-0.20.0-py2.py3-none-any.whl.metadata (1.0 kB)
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.10/dist-
packages (from jupyterthemes) (5.7.2)
```

```
Requirement already satisfied: notebook>=5.6.0 in
/usr/local/lib/python3.10/dist-packages (from jupyterthemes) (6.5.5)
Requirement already satisfied: ipython>=5.4.1 in /usr/local/lib/python3.10/dist-
packages (from jupyterthemes) (7.34.0)
Requirement already satisfied: matplotlib>=1.4.3 in
/usr/local/lib/python3.10/dist-packages (from jupyterthemes) (3.7.1)
Collecting lesscpy>=0.11.2 (from jupyterthemes)
  Downloading lesscpy-0.15.1-py2.py3-none-any.whl.metadata (6.0 kB)
Requirement already satisfied: setuptools>=18.5 in
/usr/local/lib/python3.10/dist-packages (from ipython>=5.4.1->jupyterthemes)
(71.0.4)
Collecting jedi>=0.16 (from ipython>=5.4.1->jupyterthemes)
  Downloading jedi-0.19.1-py2.py3-none-any.whl.metadata (22 kB)
Requirement already satisfied: decorator in /usr/local/lib/python3.10/dist-
packages (from ipython>=5.4.1->jupyterthemes) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-
packages (from ipython>=5.4.1->jupyterthemes) (0.7.5)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.10/dist-
packages (from ipython>=5.4.1->jupyterthemes) (5.7.1)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from ipython>=5.4.1->jupyterthemes)
(3.0.48)
Requirement already satisfied: pygments in /usr/local/lib/python3.10/dist-
packages (from ipython>=5.4.1->jupyterthemes) (2.18.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.10/dist-
packages (from ipython>=5.4.1->jupyterthemes) (0.2.0)
Requirement already satisfied: matplotlib-inline in
/usr/local/lib/python3.10/dist-packages (from ipython>=5.4.1->jupyterthemes)
(0.1.7)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.10/dist-
packages (from ipython>=5.4.1->jupyterthemes) (4.9.0)
Collecting ply (from lesscpy>=0.11.2->jupyterthemes)
  Downloading ply-3.11-py2.py3-none-any.whl.metadata (844 bytes)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib>=1.4.3->jupyterthemes)
(1.3.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-
packages (from matplotlib>=1.4.3->jupyterthemes) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib>=1.4.3->jupyterthemes)
(4.54.1)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib>=1.4.3->jupyterthemes)
(1.4.7)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.10/dist-
packages (from matplotlib>=1.4.3->jupyterthemes) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib>=1.4.3->jupyterthemes)
```

```
(24.1)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-
packages (from matplotlib>=1.4.3->jupyterthemes) (10.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib>=1.4.3->jupyterthemes)
(3.1.4)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.10/dist-packages (from matplotlib>=1.4.3->jupyterthemes)
(2.8.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from notebook>=5.6.0->jupyterthemes) (3.1.4)
Requirement already satisfied: tornado>=6.1 in /usr/local/lib/python3.10/dist-
packages (from notebook>=5.6.0->jupyterthemes) (6.3.3)
Requirement already satisfied: pyzmq<25,>=17 in /usr/local/lib/python3.10/dist-
packages (from notebook>=5.6.0->jupyterthemes) (24.0.1)
Requirement already satisfied: argon2-cffi in /usr/local/lib/python3.10/dist-
packages (from notebook>=5.6.0->jupyterthemes) (23.1.0)
Requirement already satisfied: jupyter-client<8,>=5.3.4 in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(6.1.12)
Requirement already satisfied: ipython-genutils in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(0.2.0)
Requirement already satisfied: nbformat in /usr/local/lib/python3.10/dist-
packages (from notebook>=5.6.0->jupyterthemes) (5.10.4)
Requirement already satisfied: nbconvert>=5 in /usr/local/lib/python3.10/dist-
packages (from notebook>=5.6.0->jupyterthemes) (6.5.4)
Requirement already satisfied: nest-asyncio>=1.5 in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(1.6.0)
Requirement already satisfied: ipykernel in /usr/local/lib/python3.10/dist-
packages (from notebook>=5.6.0->jupyterthemes) (5.5.6)
Requirement already satisfied: Send2Trash>=1.8.0 in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(1.8.3)
Requirement already satisfied: terminado>=0.8.3 in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(0.18.1)
Requirement already satisfied: prometheus-client in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(0.21.0)
Requirement already satisfied: nbclassic>=0.4.7 in
/usr/local/lib/python3.10/dist-packages (from notebook>=5.6.0->jupyterthemes)
(1.1.0)
Requirement already satisfied: platformdirs>=2.5 in
/usr/local/lib/python3.10/dist-packages (from jupyter-core->jupyterthemes)
(4.3.6)
Requirement already satisfied: parso<0.9.0,>=0.8.3 in
```

```
/usr/local/lib/python3.10/dist-packages (from
jedi>=0.16->ipython>=5.4.1->jupyterthemes) (0.8.4)
Requirement already satisfied: notebook-shim>=0.2.3 in
/usr/local/lib/python3.10/dist-packages (from
nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (0.2.4)
Requirement already satisfied: lxml in /usr/local/lib/python3.10/dist-packages
(from nbconvert>=5->notebook>=5.6.0->jupyterthemes) (4.9.4)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook>=5.6.0->jupyterthemes) (4.12.3)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages
(from nbconvert>=5->notebook>=5.6.0->jupyterthemes) (6.1.0)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook>=5.6.0->jupyterthemes) (0.7.1)
Requirement already satisfied: entrypoints>=0.2.2 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=5.6.0->jupyterthemes) (0.4)
Requirement already satisfied: jupyterlab-pygments in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=5.6.0->jupyterthemes) (0.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=5.6.0->jupyterthemes) (3.0.1)
Requirement already satisfied: mistune<2,>=0.8.1 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=5.6.0->jupyterthemes) (0.8.4)
Requirement already satisfied: nbclient>=0.5.0 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=5.6.0->jupyterthemes) (0.10.0)
Requirement already satisfied: pandocfilters>=1.4.1 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=5.6.0->jupyterthemes) (1.5.1)
Requirement already satisfied: tinycss2 in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook>=5.6.0->jupyterthemes) (1.3.0)
Requirement already satisfied: fastjsonschema>=2.15 in
/usr/local/lib/python3.10/dist-packages (from
nbformat->notebook>=5.6.0->jupyterthemes) (2.20.0)
Requirement already satisfied: jsonschema>=2.6 in
/usr/local/lib/python3.10/dist-packages (from
nbformat->notebook>=5.6.0->jupyterthemes) (4.23.0)
Requirement already satisfied: ptyprocess>=0.5 in
/usr/local/lib/python3.10/dist-packages (from
pexpect>4.3->ipython>=5.4.1->jupyterthemes) (0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.10/dist-
packages (from prompt-
toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython>=5.4.1->jupyterthemes) (0.2.13)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.7->matplotlib>=1.4.3->jupyterthemes) (1.16.0)
Requirement already satisfied: argon2-cffi-bindings in
```

```
/usr/local/lib/python3.10/dist-packages (from
argon2-cffi->notebook>=5.6.0->jupyterthemes) (21.2.0)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=2.6->nbformat->notebook>=5.6.0->jupyterthemes)
(24.2.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.10/dist-packages (from
jsonschema>=2.6->nbformat->notebook>=5.6.0->jupyterthemes) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.10/dist-packages (from
jsonschema>=2.6->nbformat->notebook>=5.6.0->jupyterthemes) (0.35.1)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=2.6->nbformat->notebook>=5.6.0->jupyterthemes)
(0.20.0)
Requirement already satisfied: jupyter-server<3,>=1.8 in
/usr/local/lib/python3.10/dist-packages (from notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (1.24.0)
Requirement already satisfied: cffi>=1.0.1 in /usr/local/lib/python3.10/dist-
packages (from argon2-cffi-
bindings->argon2-cffi->notebook>=5.6.0->jupyterthemes) (1.17.1)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-
packages (from beautifulsoup4->nbconvert>=5->notebook>=5.6.0->jupyterthemes)
(2.6)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-
packages (from bleach->nbconvert>=5->notebook>=5.6.0->jupyterthemes) (0.5.1)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-
packages (from cffi>=1.0.1->argon2-cffi-
bindings->argon2-cffi->notebook>=5.6.0->jupyterthemes) (2.22)
Requirement already satisfied: anyio<4,>=3.1.0 in
/usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (3.7.1)
Requirement already satisfied: websocket-client in
/usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (1.8.0)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (3.10)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (1.3.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook>=5.6.0->jupyterthemes) (1.2.2)
Downloading jupyterthemes-0.20.0-py2.py3-none-any.whl (7.0 MB)
                         7.0/7.0 MB
26.6 MB/s eta 0:00:00
Downloading lesscpy-0.15.1-py2.py3-none-any.whl (46 kB)
                         46.7/46.7 kB
```

```
3.8 MB/s eta 0:00:00
Downloading jedi-0.19.1-py2.py3-none-any.whl (1.6 MB)
                         1.6/1.6 MB
55.1 MB/s eta 0:00:00
Downloading ply-3.11-py2.py3-none-any.whl (49 kB)
                         49.6/49.6 kB
4.5 MB/s eta 0:00:00
Installing collected packages: ply, lesscpy, jedi, jupyterthemes
Successfully installed jedi-0.19.1 jupyterthemes-0.20.0 lesscpy-0.15.1 ply-3.11
```

[ ]:

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

[ ]:

```
cancer_df.tail(20)
```

[ ]:

```
     Age  Number of sexual partners  First sexual intercourse  \
838   35                        3.0                      18.0
839   31                        3.0                      19.0
840   24                        2.0                      16.0
841   23                        2.0                      15.0
842   36                        3.0                      16.0
843   30                        3.0                      14.0
844   26                        8.0                      15.0
845   19                        2.0                      15.0
846   35                        2.0                      17.0
847   30                        3.0                      22.0
848   31                        3.0                      18.0
849   32                        3.0                      18.0
850   19                        1.0                      14.0
851   23                        2.0                      15.0
852   43                        3.0                      17.0
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)  \
838                3.0    0.0            0.0                 0.0
839                1.0    0.0            0.0                 0.0
840                3.0    0.0            0.0                 0.0
841                0.0    0.0            0.0                 0.0
842                3.0    1.0            6.0                 0.3
843                3.0    0.0            0.0                 0.0
844                1.0    1.0            9.0                1.35
845                2.0    0.0            0.0                 0.0
846                1.0    0.0            0.0                 0.0
847                1.0    0.0            0.0                 0.0
```

```
848                1.0    0.0            0.0                 0.0
849                1.0    1.0           11.0                0.16
850                0.0    0.0            0.0                 0.0
851                2.0    0.0            0.0                 0.0
852                3.0    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  …  \
838                     1.0                             5.0  0.0  …
839                     1.0                            0.08  1.0  …
840                     1.0                             5.0  0.0  …
841                     0.0                             0.0  0.0  …
842                     1.0                             2.0  0.0  …
843                     1.0                             2.0  0.0  …
844                     1.0                             5.0  1.0  …
845                     1.0                            0.75  0.0  …
846                     0.0                             0.0  0.0  …
847                     0.0                             0.0  0.0  …
848                     1.0                             0.5  0.0  …
849                     1.0                             6.0  0.0  …
850                     0.0                             0.0  0.0  …
851                     0.0                             0.0  0.0  …
852                     1.0                             5.0  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  \
838                                ?                               ?
839                                ?                               ?
840                                ?                               ?
841                                ?                               ?
842                                ?                               ?
843                                ?                               ?
844                                ?                               ?
845                                ?                               ?
846                                ?                               ?
847                                ?                               ?
848                                ?                               ?
849                                ?                               ?
850                                ?                               ?
```

```
851                                  ?                               ?
852                                  ?                               ?
853                                  ?                               ?
854                                  ?                               ?
855                                  ?                               ?
856                                  ?                               ?
857                                  ?                               ?

    Dx:Cancer Dx:CIN Dx:HPV Dx Hinselmann Schiller Citology Biopsy
838         0      0      0  0          0        0        0      0
839         1      0      0  1          0        0        0      0
840         0      0      0  0          0        0        0      0
841         0      0      0  0          0        0        0      0
842         0      0      0  0          0        0        0      0
843         0      0      0  0          0        0        0      0
844         0      0      0  0          0        0        0      0
845         0      0      0  0          0        0        0      0
846         0      0      0  0          0        0        0      0
847         0      0      0  0          0        0        0      0
848         0      0      0  0          0        0        0      0
849         1      0      1  0          0        0        0      0
850         0      0      0  0          0        0        0      0
851         0      0      0  0          0        0        0      0
852         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

[20 rows x 36 columns]
```

1.0.2 STEP 2: EXPLORATORY DATA ANALYSIS

```
[ ]: 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
```

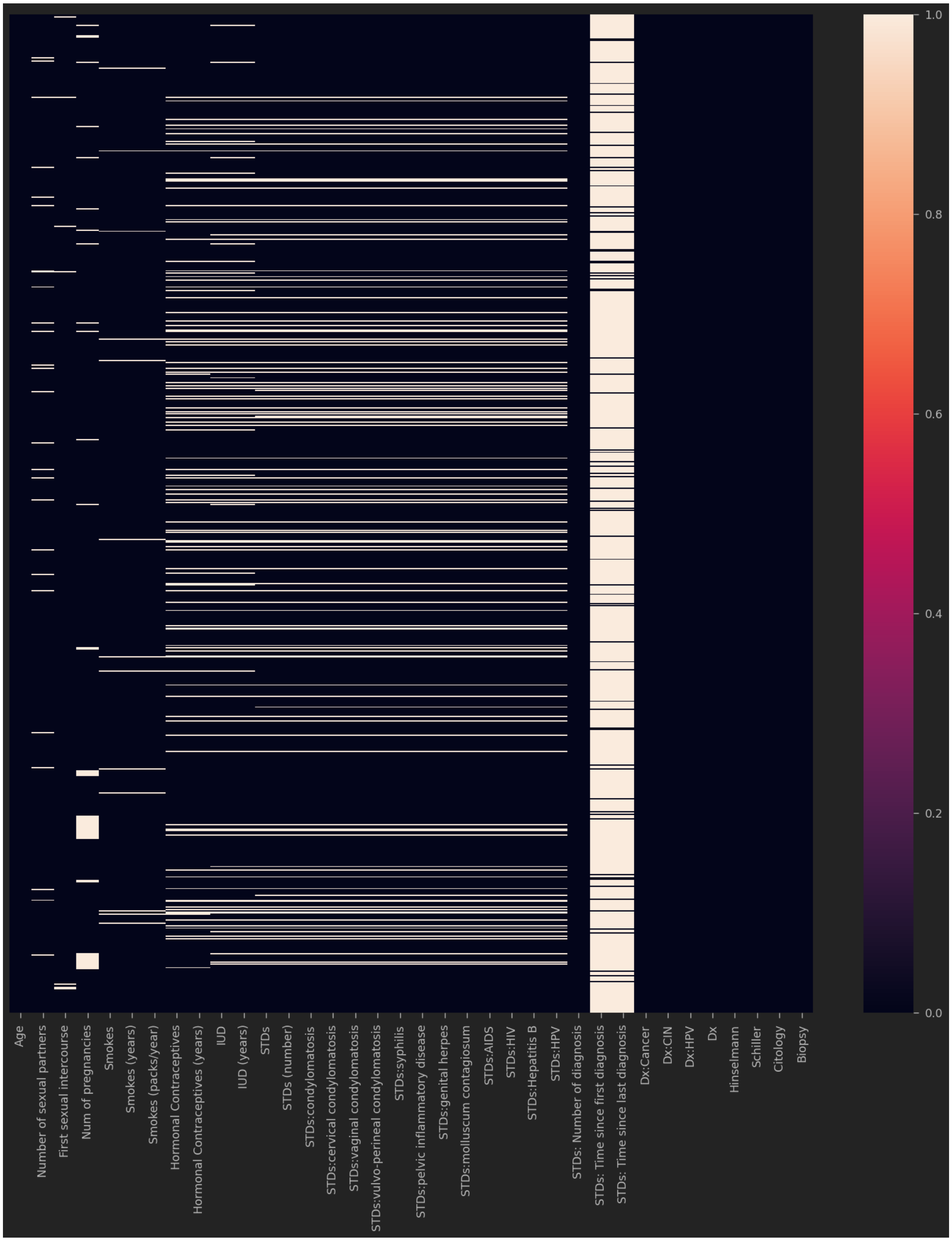
[ ]:
```
# REPLACING '?' WITH NaN

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]
```

[ ]:

```
# PLOTTING HEATMAP TO VISUALIZE THE NUMBER OF NaN'S IN TH DATA

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



[ ]:

```
# WE OBSERVE THAT THERE ARE A LOT OF NAN VALUES IN "STD'S: TIME SINCE FIRST␣
 ↪DIAGNOSIS" AND "STD'S: TIME SINCE LAST DIAGNOSIS"
# SO WE WILL DROP THESE COLUMNS
```

```
cancer_df = cancer_df.drop(['STDs: Time since first diagnosis', 'STDs: Time␣
↪since last diagnosis'], axis=1)
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]
```

[ ]:
```python
# Converting the column data types, from object to numeric in order to perform Statistical Analysis of the Data

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]
```

```
[ ]: 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
```

[ ]:

```
# REPLACING NULL/NaN values with the mean values:

cancer_df =  cancer_df.fillna(cancer_df.mean())
cancer_df
```

[ ]:

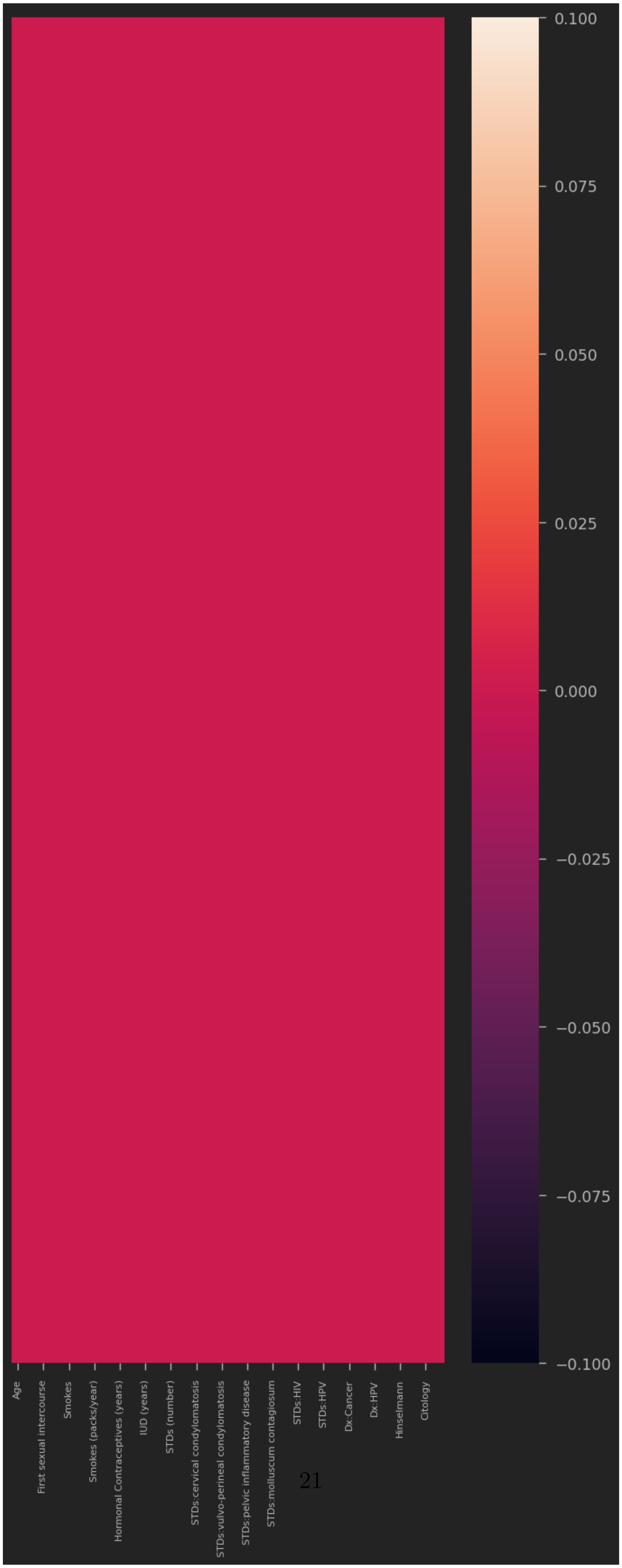
```
     Age  Number of sexual partners  First sexual intercourse  \
0     18                        4.0                   15.0000
1     15                        1.0                   14.0000
2     34                        1.0                   16.9953
3     52                        5.0                   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]
```

[ ]:

```
# PLOTTING HEATMAP AGAIN TO VISUALIZE AND CHECK OUR DATA CLEANSING

plt.figure(figsize=(8,20))
sns.heatmap(cancer_df.isnull(), yticklabels = False)
plt.xticks(rotation=90)
plt.tick_params(labelsize=8)
plt.show()
```


0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
Age
First sexual intercourse
Smokes
Smokes (packs/year)
Hormonal Contraceptives (years)
IUD (years)
STDs (number)
STDs:cervical condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:pelvic inflammatory disease
STDs:molluscum contagiosum
STDs:HIV
STDs:HPV
Dx:Cancer
Dx:HPV
Hinselmann
Citology

[ ]:

```
# THUS WE CAN SEE THAT WE HAVE NO NULL VALUES NOW
```

[ ]:

```
cancer_df.describe()
```

[ ]:

```
              Age  Number of sexual partners  First sexual intercourse  \
count  858.000000                 858.000000                858.000000
mean    26.820513                   2.527644                 16.995300
std      8.497948                   1.642267                  2.791883
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          858.000000  858.000000      858.000000           858.000000
mean             2.275561    0.145562        1.219721             0.453144
std              1.399325    0.350189        4.057885             2.209657
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               858.000000                       858.000000  858.000000
mean                  0.641333                         2.256419    0.112011
std                   0.448671                         3.519082    0.293260
min                   0.000000                         0.000000    0.000000
25%                   0.000000                         0.000000    0.000000
50%                   1.000000                         1.000000    0.000000
75%                   1.000000                         2.256419    0.000000
max                   1.000000                        30.000000    1.000000

       …    STDs:HPV  STDs: Number of diagnosis   Dx:Cancer      Dx:CIN  \
count  …  858.000000                 858.000000  858.000000  858.000000
mean   …    0.002656                   0.087413    0.020979    0.010490
std    …    0.048244                   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]
```

[ ]:
```
# THIS THE AGE RANGE FOR PEOPLE INVOLVED IN THE STUDY ARE : (13, 84)
```

1.0.3 3) PERFORMING DATA VISUALIZATION

[ ]:
```
# WE'LL TRY TO OBSERVE THE CORELATION BETWEEN DIFFERENT FEATURES IN OUR␣
 ↪DATASETS:

corr_matrix = cancer_df.corr()

corr_matrix
```

[ ]:
```
                                         Age  Number of sexual partners  \
Age                                 1.000000                   0.084896
Number of sexual partners           0.084896                   1.000000
First sexual intercourse            0.369168                  -0.147937
Num of pregnancies                  0.526137                   0.076719
Smokes                              0.055813                   0.238078
Smokes (years)                      0.217349                   0.177117
Smokes (packs/year)                 0.131180                   0.175153
Hormonal Contraceptives             0.065624                   0.006342
Hormonal Contraceptives (years)     0.277181                   0.018552
IUD                                 0.267662                   0.030005
IUD (years)                         0.205886                   0.004215
STDs                                0.010017                   0.053754
STDs (number)                      -0.015488                   0.039359
STDs:condylomatosis                -0.025012                   0.034646
STDs:cervical condylomatosis             NaN                        NaN
STDs:vaginal condylomatosis         0.006220                  -0.042924
STDs:vulvo-perineal condylomatosis -0.022614                   0.036750
STDs:syphilis                       0.010442                   0.027178
STDs:pelvic inflammatory disease    0.023216                   0.030616
STDs:genital herpes                -0.029076                  -0.031826
STDs:molluscum contagiosum         -0.000919                   0.030616
STDs:AIDS                                NaN                        NaN
STDs:HIV                           -0.002025                   0.019871
```

```
STDs:Hepatitis B                   -0.029076                  -0.011012
STDs:HPV                            0.038546                   0.013871
STDs: Number of diagnosis          -0.001606                   0.051559
Dx:Cancer                           0.110340                   0.022309
Dx:CIN                              0.061443                   0.015691
Dx:HPV                              0.101722                   0.027264
Dx                                  0.092635                   0.022982
Hinselmann                         -0.003967                  -0.039273
Schiller                            0.103283                  -0.008899
Citology                           -0.016862                   0.021839
Biopsy                              0.055956                  -0.001429

                                    First sexual intercourse  \
Age                                                 0.369168
Number of sexual partners                          -0.147937
First sexual intercourse                            1.000000
Num of pregnancies                                 -0.058223
Smokes                                             -0.123602
Smokes (years)                                     -0.058366
Smokes (packs/year)                                -0.056332
Hormonal Contraceptives                             0.018344
Hormonal Contraceptives (years)                     0.008000
IUD                                                -0.020975
IUD (years)                                        -0.024803
STDs                                               -0.013133
STDs (number)                                       0.006487
STDs:condylomatosis                                 0.026777
STDs:cervical condylomatosis                             NaN
STDs:vaginal condylomatosis                         0.071425
STDs:vulvo-perineal condylomatosis                  0.031082
STDs:syphilis                                      -0.100999
STDs:pelvic inflammatory disease                   -0.001089
STDs:genital herpes                                 0.023398
STDs:molluscum contagiosum                         -0.013332
STDs:AIDS                                                NaN
STDs:HIV                                           -0.013430
STDs:Hepatitis B                                    0.011154
STDs:HPV                                            0.033112
STDs: Number of diagnosis                          -0.013327
Dx:Cancer                                           0.067283
Dx:CIN                                             -0.032626
Dx:HPV                                              0.043966
Dx                                                  0.035750
Hinselmann                                         -0.016546
Schiller                                            0.003493
Citology                                           -0.010971
Biopsy                                              0.007262
```

```
                                    Num of pregnancies    Smokes  \
Age                                           0.526137  0.055813
Number of sexual partners                     0.076719  0.238078
First sexual intercourse                     -0.058223 -0.123602
Num of pregnancies                            1.000000  0.080768
Smokes                                        0.080768  1.000000
Smokes (years)                                0.174912  0.723128
Smokes (packs/year)                           0.097044  0.493361
Hormonal Contraceptives                       0.142858 -0.002165
Hormonal Contraceptives (years)               0.207839  0.044157
IUD                                           0.198550 -0.051184
IUD (years)                                   0.143642 -0.032996
STDs                                          0.044250  0.116676
STDs (number)                                 0.001706  0.105811
STDs:condylomatosis                          -0.037999  0.059919
STDs:cervical condylomatosis                       NaN       NaN
STDs:vaginal condylomatosis                  -0.003166  0.069631
STDs:vulvo-perineal condylomatosis           -0.037204  0.062775
STDs:syphilis                                 0.141728  0.082684
STDs:pelvic inflammatory disease             -0.056542 -0.014059
STDs:genital herpes                          -0.032114 -0.014059
STDs:molluscum contagiosum                    0.041168 -0.014059
STDs:AIDS                                          NaN       NaN
STDs:HIV                                      0.009384  0.059412
STDs:Hepatitis B                             -0.032114  0.083551
STDs:HPV                                     -0.028162  0.049171
STDs: Number of diagnosis                     0.033514  0.095433
Dx:Cancer                                     0.035123 -0.011027
Dx:CIN                                        0.007344 -0.042822
Dx:HPV                                        0.046753  0.012210
Dx                                            0.019025 -0.067614
Hinselmann                                    0.038685  0.034527
Schiller                                      0.087687  0.053613
Citology                                     -0.029656 -0.003913
Biopsy                                        0.043460  0.029091

                                    Smokes (years)  Smokes (packs/year)  \
Age                                       0.217349             0.131180
Number of sexual partners                 0.177117             0.175153
First sexual intercourse                 -0.058366            -0.056332
Num of pregnancies                        0.174912             0.097044
Smokes                                    0.723128             0.493361
Smokes (years)                            1.000000             0.724116
Smokes (packs/year)                       0.724116             1.000000
Hormonal Contraceptives                  -0.011002             0.005880
Hormonal Contraceptives (years)           0.048899             0.040112
```

```
IUD                                       0.027562             0.007891
IUD (years)                               0.037900             0.015912
STDs                                      0.091611             0.029372
STDs (number)                             0.091313             0.030780
STDs:condylomatosis                       0.045397             0.007917
STDs:cervical condylomatosis                   NaN                  NaN
STDs:vaginal condylomatosis               0.114332             0.041412
STDs:vulvo-perineal condylomatosis        0.047511             0.009130
STDs:syphilis                             0.015393            -0.003277
STDs:pelvic inflammatory disease         -0.010337            -0.007180
STDs:genital herpes                      -0.010337            -0.007180
STDs:molluscum contagiosum               -0.010337            -0.007180
STDs:AIDS                                      NaN                  NaN
STDs:HIV                                  0.090636             0.054577
STDs:Hepatitis B                          0.099170             0.101105
STDs:HPV                                  0.050935            -0.008410
STDs: Number of diagnosis                 0.081676             0.032186
Dx:Cancer                                 0.054674             0.108476
Dx:CIN                                   -0.030966            -0.021127
Dx:HPV                                    0.057214             0.110366
Dx                                       -0.048894            -0.033358
Hinselmann                                0.071232             0.026662
Schiller                                  0.094640             0.017954
Citology                                 -0.006750             0.004613
Biopsy                                    0.061484             0.024657

                                    Hormonal Contraceptives  \
Age                                                0.065624
Number of sexual partners                          0.006342
First sexual intercourse                           0.018344
Num of pregnancies                                 0.142858
Smokes                                            -0.002165
Smokes (years)                                    -0.011002
Smokes (packs/year)                                0.005880
Hormonal Contraceptives                            1.000000
Hormonal Contraceptives (years)                    0.448574
IUD                                                0.033729
IUD (years)                                       -0.033752
STDs                                              -0.032105
STDs (number)                                     -0.038088
STDs:condylomatosis                               -0.009284
STDs:cervical condylomatosis                            NaN
STDs:vaginal condylomatosis                       -0.059222
STDs:vulvo-perineal condylomatosis                -0.013714
STDs:syphilis                                     -0.003624
STDs:pelvic inflammatory disease                   0.027587
STDs:genital herpes                                0.027587
```

```
STDs:molluscum contagiosum                        -0.048598
STDs:AIDS                                               NaN
STDs:HIV                                          -0.076278
STDs:Hepatitis B                                  -0.048598
STDs:HPV                                           0.039040
STDs: Number of diagnosis                         -0.050660
Dx:Cancer                                          0.026407
Dx:CIN                                            -0.003334
Dx:HPV                                             0.038038
Dx                                                -0.001723
Hinselmann                                         0.033551
Schiller                                          -0.004247
Citology                                          -0.011030
Biopsy                                             0.007711

                                    Hormonal Contraceptives (years)       IUD  \
Age                                                        0.277181  0.267662
Number of sexual partners                                  0.018552  0.030005
First sexual intercourse                                   0.008000 -0.020975
Num of pregnancies                                         0.207839  0.198550
Smokes                                                     0.044157 -0.051184
Smokes (years)                                             0.048899  0.027562
Smokes (packs/year)                                        0.040112  0.007891
Hormonal Contraceptives                                    0.448574  0.033729
Hormonal Contraceptives (years)                            1.000000  0.094953
IUD                                                        0.094953  1.000000
IUD (years)                                                0.000455  0.746478
STDs                                                       0.000829  0.053859
STDs (number)                                             -0.006468  0.053146
STDs:condylomatosis                                        0.007752  0.077262
STDs:cervical condylomatosis                                    NaN       NaN
STDs:vaginal condylomatosis                               -0.038207  0.032093
STDs:vulvo-perineal condylomatosis                         0.009685  0.061867
STDs:syphilis                                              0.003897 -0.022311
STDs:pelvic inflammatory disease                          -0.014209 -0.013125
STDs:genital herpes                                       -0.019065 -0.013125
STDs:molluscum contagiosum                                -0.021494 -0.013125
STDs:AIDS                                                       NaN       NaN
STDs:HIV                                                  -0.035472  0.008590
STDs:Hepatitis B                                          -0.021494 -0.013125
STDs:HPV                                                   0.052059 -0.018574
STDs: Number of diagnosis                                 -0.037219  0.029871
Dx:Cancer                                                  0.054627  0.110541
Dx:CIN                                                     0.003086  0.051833
Dx:HPV                                                     0.061394  0.058154
Dx                                                        -0.012865  0.138905
Hinselmann                                                 0.038825  0.044059
```

```
Schiller                                                   0.078707  0.084074
Citology                                                   0.074324  0.007348
Biopsy                                                     0.078995  0.051554

                                    …  STDs:HPV  STDs: Number of diagnosis  \
Age                                 …  0.038546                  -0.001606
Number of sexual partners           …  0.013871                   0.051559
First sexual intercourse            …  0.033112                  -0.013327
Num of pregnancies                  … -0.028162                   0.033514
Smokes                              …  0.049171                   0.095433
Smokes (years)                      …  0.050935                   0.081676
Smokes (packs/year)                 … -0.008410                   0.032186
Hormonal Contraceptives             …  0.039040                  -0.050660
Hormonal Contraceptives (years)     …  0.052059                  -0.037219
IUD                                 … -0.018574                   0.029871
IUD (years)                         … -0.013865                   0.007601
STDs                                …  0.150734                   0.901364
STDs (number)                       …  0.075657                   0.891990
STDs:condylomatosis                 … -0.012856                   0.694953
STDs:cervical condylomatosis        …       NaN                        NaN
STDs:vaginal condylomatosis         … -0.003771                   0.203864
STDs:vulvo-perineal condylomatosis  … -0.012700                   0.686526
STDs:syphilis                       … -0.008076                   0.409625
STDs:pelvic inflammatory disease    … -0.001882                   0.101729
STDs:genital herpes                 … -0.001882                   0.101729
STDs:molluscum contagiosum          … -0.001882                   0.101729
STDs:AIDS                           …       NaN                        NaN
STDs:HIV                            … -0.008076                   0.544306
STDs:Hepatitis B                    … -0.001882                   0.101729
STDs:HPV                            …  1.000000                   0.064018
STDs: Number of diagnosis           …  0.064018                   1.000000
Dx:Cancer                           …  0.329270                  -0.015423
Dx:CIN                              … -0.005041                   0.008070
Dx:HPV                              …  0.329270                  -0.015423
Dx                                  …  0.137639                  -0.002289
Hinselmann                          … -0.011360                   0.076787
Schiller                            … -0.016695                   0.130873
Citology                            … -0.011934                   0.055114
Biopsy                              … -0.013892                   0.097449

                                    Dx:Cancer    Dx:CIN    Dx:HPV        Dx  \
Age                                  0.110340  0.061443  0.101722  0.092635
Number of sexual partners            0.022309  0.015691  0.027264  0.022982
First sexual intercourse             0.067283 -0.032626  0.043966  0.035750
Num of pregnancies                   0.035123  0.007344  0.046753  0.019025
Smokes                              -0.011027 -0.042822  0.012210 -0.067614
Smokes (years)                       0.054674 -0.030966  0.057214 -0.048894
Smokes (packs/year)                  0.108476 -0.021127  0.110366 -0.033358
Hormonal Contraceptives              0.026407 -0.003334  0.038038 -0.001723
Hormonal Contraceptives (years)      0.054627  0.003086  0.061394 -0.012865
IUD                                  0.110541  0.051833  0.058154  0.138905
IUD (years)                          0.097947  0.014715  0.032666  0.102287
STDs                                 0.003160  0.006403  0.003160 -0.010169
STDs (number)                       -0.018228 -0.008980 -0.018228 -0.027707
STDs:condylomatosis                 -0.038927 -0.024337 -0.038927 -0.043230
STDs:cervical condylomatosis              NaN       NaN       NaN       NaN
STDs:vaginal condylomatosis         -0.011419 -0.007139 -0.011419 -0.012682
STDs:vulvo-perineal condylomatosis  -0.038455 -0.024042 -0.038455 -0.042706
STDs:syphilis                       -0.024453 -0.015288 -0.024453 -0.027157
STDs:pelvic inflammatory disease    -0.005698 -0.003562 -0.005698 -0.006328
STDs:genital herpes                 -0.005698 -0.003562 -0.005698 -0.006328
STDs:molluscum contagiosum          -0.005698 -0.003562 -0.005698 -0.006328
STDs:AIDS                                 NaN       NaN       NaN       NaN
STDs:HIV                            -0.024453  0.064656 -0.024453  0.022237
STDs:Hepatitis B                    -0.005698 -0.003562 -0.005698 -0.006328
STDs:HPV                             0.329270 -0.005041  0.329270  0.137639
STDs: Number of diagnosis           -0.015423  0.008070 -0.015423 -0.002289
Dx:Cancer                            1.000000 -0.015072  0.886508  0.665647
Dx:CIN                              -0.015072  1.000000 -0.015072  0.606939
Dx:HPV                               0.886508 -0.015072  1.000000  0.616327
Dx                                   0.665647  0.606939  0.616327  1.000000
Hinselmann                           0.134264 -0.021233  0.134264  0.072215
Schiller                             0.157812  0.009119  0.157812  0.098952
Citology                             0.113446 -0.023938  0.113446  0.088740
Biopsy                               0.160905  0.113172  0.160905  0.157607

                                    Hinselmann  Schiller  Citology    Biopsy
Age                                  -0.003967  0.103283 -0.016862  0.055956
Number of sexual partners            -0.039273 -0.008899  0.021839 -0.001429
First sexual intercourse             -0.016546  0.003493 -0.010971  0.007262
Num of pregnancies                    0.038685  0.087687 -0.029656  0.043460
Smokes                                0.034527  0.053613 -0.003913  0.029091
Smokes (years)                        0.071232  0.094640 -0.006750  0.061484
Smokes (packs/year)                   0.026662  0.017954  0.004613  0.024657
Hormonal Contraceptives               0.033551 -0.004247 -0.011030  0.007711
Hormonal Contraceptives (years)       0.038825  0.078707  0.074324  0.078995
IUD                                   0.044059  0.084074  0.007348  0.051554
IUD (years)                           0.007858  0.077865  0.002615  0.032250
STDs                                  0.047780  0.106169  0.049669  0.106737
STDs (number)                         0.065155  0.119203  0.057831  0.096218
STDs:condylomatosis                   0.052417  0.108344  0.062623  0.084520
STDs:cervical condylomatosis               NaN       NaN       NaN       NaN
STDs:vaginal condylomatosis          -0.016087 -0.023642 -0.016900 -0.019673
STDs:vulvo-perineal condylomatosis    0.054245  0.111371  0.064626  0.086977
```

```
STDs:syphilis                          0.006726  0.007398 -0.036190 -0.042128
STDs:pelvic inflammatory disease      -0.008027 -0.011797 -0.008433 -0.009817
STDs:genital herpes                   -0.008027 -0.011797 -0.008433  0.129657
STDs:molluscum contagiosum            -0.008027 -0.011797 -0.008433 -0.009817
STDs:AIDS                                   NaN       NaN       NaN       NaN
STDs:HIV                               0.089074  0.123448  0.074586  0.124133
STDs:Hepatitis B                      -0.008027 -0.011797 -0.008433 -0.009817
STDs:HPV                              -0.011360 -0.016695 -0.011934 -0.013892
STDs: Number of diagnosis              0.076787  0.130873  0.055114  0.097449
Dx:Cancer                              0.134264  0.157812  0.113446  0.160905
Dx:CIN                                -0.021233  0.009119 -0.023938  0.113172
Dx:HPV                                 0.134264  0.157812  0.113446  0.160905
Dx                                     0.072215  0.098952  0.088740  0.157607
Hinselmann                             1.000000  0.650249  0.192467  0.547417
Schiller                               0.650249  1.000000  0.361486  0.733204
Citology                               0.192467  0.361486  1.000000  0.327466
Biopsy                                 0.547417  0.733204  0.327466  1.000000

[34 rows x 34 columns]
```

[ ]:

```python
# PLOTTING THE HEATMAP FOR CORRELATION MATRIX

plt.figure(figsize = (30,30))
sns.heatmap(corr_matrix, annot=True)
plt.xticks(rotation=90)
plt.yticks(rotation=360)
plt.tick_params(labelsize=8)
plt.show()
```

[ ]:

```python
# VISUALIZING THE WHOLE DATAFRAME BY PLOTTING HISTOGRAM
cancer_df.hist(bins = 10, figsize = (30,30), color='blue')
plt.show()
```

1.0.4 4) PREPARING DATA BEFORE TRAINING

[ ]:

```python
# WE SELECT BIOPSY DATA AS OUR TARGET VALUES:

target_df = cancer_df['Biopsy']
input_df = cancer_df.drop(['Biopsy'], axis=1)
```

[ ]:

```python
X = np.array(input_df).astype('float32')
y = np.array(target_df).astype('float32')

y = y.reshape(-1,1)
```

[ ]:

```python
from sklearn.preprocessing import StandardScaler, MinMaxScaler

scaler = StandardScaler()
X = scaler.fit_transform(X)
```

[ ]:

```python
X
```

[ ]:

```python
# SPLITTING DATA INTO TRAIN AND TEST DATASETS
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)
X_test, X_val, y_test, y_val = train_test_split(X_test, y_test, test_size = 0.5)
```

1.0.5 5) TRAINING AND EVALUATING XGBOOST CLASSIFIER

[ ]:

```
!pip install --upgrade pip
!pip install seaborn
!pip install xgboost
```

[ ]:

```python
import xgboost as xgb

model = xgb.XGBClassifier(learning_rate = 0.1, max_depth = 50, n_estimators =␣
  ↪100)

model.fit(X_train, y_train)
```

TESTING OUT RESULTS OF OUR MODEL

[ ]:

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

result_train
```

[ ]:

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

result_test
```

[ ]:

```python
y_predict = model.predict(X_test)
```

[ ]:

```python
from sklearn.metrics import confusion_matrix, classification_report

print(classification_report(y_test, y_predict))
```

PLOTTING A CONFUSION MATRIX

[ ]:

```python
cm = confusion_matrix(y_predict, y_test)

sns.heatmap(cm, annot = True)
```

```
plt.show()
```

[ ]:

```
```