Hands_On_Activity_11_2_(CALINGO) (2)

April 28, 2024

[ ]:
```
pip install ucimlrepo
```

```
Collecting ucimlrepo
  Downloading ucimlrepo-0.0.6-py3-none-any.whl (8.0 kB)
Installing collected packages: ucimlrepo
Successfully installed ucimlrepo-0.0.6
```

[ ]:
```
from ucimlrepo import fetch_ucirepo

# fetch dataset
cervical_cancer_risk_factors = fetch_ucirepo(id=383)

# data (as pandas dataframes)
X = cervical_cancer_risk_factors.data.features
y = cervical_cancer_risk_factors.data.targets

# metadata
print(cervical_cancer_risk_factors.metadata)

# variable information
print(cervical_cancer_risk_factors.variables)
```

```
{'uci_id': 383, 'name': 'Cervical Cancer (Risk Factors)', 'repository_url':
'https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors',
'data_url': 'https://archive.ics.uci.edu/static/public/383/data.csv',
'abstract': 'This dataset focuses on the prediction of indicators/diagnosis of
cervical cancer. The features cover demographic information, habits, and
historic medical records.', 'area': 'Health and Medicine', 'tasks':
['Classification'], 'characteristics': ['Multivariate'], 'num_instances': 858,
'num_features': 36, 'feature_types': ['Integer', 'Real'], 'demographics':
['Age', 'Other'], 'target_col': None, 'index_col': None, 'has_missing_values':
'yes', 'missing_values_symbol': 'NaN', 'year_of_dataset_creation': 2017,
'last_updated': 'Sun Mar 10 2024', 'dataset_doi': '10.24432/C5Z310', 'creators':
['Kelwin Fernandes', 'Jaime Cardoso', 'Jessica Fernandes'], 'intro_paper':
{'title': 'Transfer Learning with Partial Observability Applied to Cervical
Cancer Screening', 'authors': 'Kelwin Fernandes, Jaime S. Cardoso, Jessica C.
Fernandes', 'published_in': 'Iberian Conference on Pattern Recognition and Image
Analysis', 'year': 2017, 'url': 'https://www.semanticscholar.org/paper/Transfer-
```

Learning-with-Partial-Observability-to-Fernandes-
Cardoso/1c02438ba4dfa775399ba414508e9cd335b69012', 'doi': None},
'additional_info': {'summary': "The dataset was collected at 'Hospital
Universitario de Caracas' in Caracas, Venezuela. The dataset comprises
demographic information, habits, and historic medical records of 858 patients.
Several patients decided not to answer some of the questions because of privacy
concerns (missing values).", 'purpose': None, 'funded_by': None,
'instances_represent': None, 'recommended_data_splits': None, 'sensitive_data':
None, 'preprocessing_description': None, 'variable_info': '(int) Age\r\n(int)
Number of sexual partners\r\n(int) First sexual intercourse (age)\r\n(int) Num
of pregnancies\r\n(bool) Smokes\r\n(bool) Smokes (years)\r\n(bool) Smokes
(packs/year)\r\n(bool) Hormonal Contraceptives\r\n(int) Hormonal Contraceptives
(years)\r\n(bool) IUD\r\n(int) IUD (years)\r\n(bool) STDs\r\n(int) STDs
(number)\r\n(bool) STDs:condylomatosis\r\n(bool) STDs:cervical
condylomatosis\r\n(bool) STDs:vaginal condylomatosis\r\n(bool) STDs:vulvo-
perineal condylomatosis\r\n(bool) STDs:syphilis\r\n(bool) STDs:pelvic
inflammatory disease\r\n(bool) STDs:genital herpes\r\n(bool) STDs:molluscum
contagiosum\r\n(bool) STDs:AIDS\r\n(bool) STDs:HIV\r\n(bool) STDs:Hepatitis
B\r\n(bool) STDs:HPV\r\n(int) STDs: Number of diagnosis\r\n(int) STDs: Time
since first diagnosis\r\n(int) STDs: Time since last diagnosis\r\n(bool)
Dx:Cancer\r\n(bool) Dx:CIN\r\n(bool) Dx:HPV\r\n(bool) Dx\r\n(bool) Hinselmann:
target variable\r\n(bool) Schiller: target variable\r\n(bool) Cytology: target
variable\r\n(bool) Biopsy: target variable', 'citation': None}}

	name	role	type	demographic
0	Age	Feature	Integer	Age
1	Number of sexual partners	Feature	Continuous	Other
2	First sexual intercourse	Feature	Continuous	None
3	Num of pregnancies	Feature	Continuous	None
4	Smokes	Feature	Continuous	None
5	Smokes (years)	Feature	Continuous	None
6	Smokes (packs/year)	Feature	Continuous	None
7	Hormonal Contraceptives	Feature	Continuous	None
8	Hormonal Contraceptives (years)	Feature	Continuous	None
9	IUD	Feature	Continuous	None
10	IUD (years)	Feature	Continuous	None
11	STDs	Feature	Continuous	None
12	STDs (number)	Feature	Continuous	None
13	STDs:condylomatosis	Feature	Continuous	None
14	STDs:cervical condylomatosis	Feature	Continuous	None
15	STDs:vaginal condylomatosis	Feature	Continuous	None
16	STDs:vulvo-perineal condylomatosis	Feature	Continuous	None
17	STDs:syphilis	Feature	Continuous	None
18	STDs:pelvic inflammatory disease	Feature	Continuous	None
19	STDs:genital herpes	Feature	Continuous	None
20	STDs:molluscum contagiosum	Feature	Continuous	None
21	STDs:AIDS	Feature	Continuous	None
22	STDs:HIV	Feature	Continuous	None
23	STDs:Hepatitis B	Feature	Continuous	None

24	STDs:HPV	Feature	Continuous	None	
25	STDs: Number of diagnosis	Feature	Integer	None	
26	STDs: Time since first diagnosis	Feature	Continuous	None	
27	STDs: Time since last diagnosis	Feature	Continuous	None	
28	Dx:Cancer	Feature	Integer	None	
29	Dx:CIN	Feature	Integer	None	
30	Dx:HPV	Feature	Integer	None	
31	Dx	Feature	Integer	None	
32	Hinselmann	Feature	Integer	None	
33	Schiller	Feature	Integer	None	
34	Citology	Feature	Integer	None	
35	Biopsy	Feature	Integer	None	

	description	units	missing_values
0	None	None	no
1	None	None	yes
2	None	None	yes
3	None	None	yes
4	None	None	yes
5	None	None	yes
6	None	None	yes
7	None	None	yes
8	None	None	yes
9	None	None	yes
10	None	None	yes
11	None	None	yes
12	None	None	yes
13	None	None	yes
14	None	None	yes
15	None	None	yes
16	None	None	yes
17	None	None	yes
18	None	None	yes
19	None	None	yes
20	None	None	yes
21	None	None	yes
22	None	None	yes
23	None	None	yes
24	None	None	yes
25	None	None	no
26	None	None	yes
27	None	None	yes
28	None	None	no
29	None	None	no
30	None	None	no
31	None	None	no
32	None	None	no
33	None	None	no

```
34        None  None             no
35        None  None             no
```

[ ]:
```
X
```

[ ]:
```
     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                             0.00  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: 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]
```

[ ]:
```
y
```

Since Y returns no value, we will not concatenate our data

[ ]:
```
!pip install hvplot
```

```
Collecting hvplot
  Downloading hvplot-0.9.2-py2.py3-none-any.whl (1.8 MB)
                             1.8/1.8 MB
8.0 MB/s eta 0:00:00
Requirement already satisfied: bokeh>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from hvplot) (3.3.4)
Requirement already satisfied: colorcet>=2 in /usr/local/lib/python3.10/dist-
packages (from hvplot) (3.1.0)
Requirement already satisfied: holoviews>=1.11.0 in
/usr/local/lib/python3.10/dist-packages (from hvplot) (1.17.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages
(from hvplot) (2.0.3)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.10/dist-
packages (from hvplot) (1.25.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
```

```
packages (from hvplot) (24.0)
Requirement already satisfied: panel>=0.11.0 in /usr/local/lib/python3.10/dist-
packages (from hvplot) (1.3.8)
Requirement already satisfied: param<3.0,>=1.12.0 in
/usr/local/lib/python3.10/dist-packages (from hvplot) (2.1.0)
Requirement already satisfied: Jinja2>=2.9 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (3.1.3)
Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (1.2.1)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (6.0.1)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-
packages (from bokeh>=1.0.0->hvplot) (6.3.3)
Requirement already satisfied: xyzservices>=2021.09.1 in
/usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (2024.4.0)
Requirement already satisfied: pyviz-comms>=0.7.4 in
/usr/local/lib/python3.10/dist-packages (from holoviews>=1.11.0->hvplot) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->hvplot) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->hvplot) (2024.1)
Requirement already satisfied: markdown in /usr/local/lib/python3.10/dist-
packages (from panel>=0.11.0->hvplot) (3.6)
Requirement already satisfied: markdown-it-py in /usr/local/lib/python3.10/dist-
packages (from panel>=0.11.0->hvplot) (3.0.0)
Requirement already satisfied: linkify-it-py in /usr/local/lib/python3.10/dist-
packages (from panel>=0.11.0->hvplot) (2.0.3)
Requirement already satisfied: mdit-py-plugins in
/usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (0.4.0)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from panel>=0.11.0->hvplot) (2.31.0)
Requirement already satisfied: tqdm>=4.48.0 in /usr/local/lib/python3.10/dist-
packages (from panel>=0.11.0->hvplot) (4.66.2)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages
(from panel>=0.11.0->hvplot) (6.1.0)
Requirement already satisfied: typing-extensions in
/usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.11.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from Jinja2>=2.9->bokeh>=1.0.0->hvplot)
(2.1.5)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.2->pandas->hvplot) (1.16.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-
packages (from bleach->panel>=0.11.0->hvplot) (0.5.1)
```

```
Requirement already satisfied: uc-micro-py in /usr/local/lib/python3.10/dist-
packages (from linkify-it-py->panel>=0.11.0->hvplot) (1.0.3)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-
packages (from markdown-it-py->panel>=0.11.0->hvplot) (0.1.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->panel>=0.11.0->hvplot) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot)
(2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot)
(2024.2.2)
Installing collected packages: hvplot
Successfully installed hvplot-0.9.2
```

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

from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
```

[145]:
```
data = X
data
```

[145]:

	Age	Number of sexual partners	First sexual intercourse	\
0	18	4.0	15.0	
1	15	1.0	14.0	
2	34	1.0	15.0	
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                             0.00  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: Time since first diagnosis  STDs: Time since last diagnosis  \
0                                 1.0                               1.0
1                                 1.0                               1.0
2                                 1.0                               1.0
3                                 1.0                               1.0
4                                 1.0                               1.0
..                                …                               …
853                               1.0                               1.0
854                               1.0                               1.0
855                               1.0                               1.0
856                               1.0                               1.0
857                               1.0                               1.0

     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]
```

[146]:
```
data.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    float64
 2   First sexual intercourse            858 non-null    float64
 3   Num of pregnancies                  858 non-null    float64
 4   Smokes                              858 non-null    float64
 5   Smokes (years)                      858 non-null    float64
 6   Smokes (packs/year)                 858 non-null    float64
 7   Hormonal Contraceptives             858 non-null    float64
 8   Hormonal Contraceptives (years)     858 non-null    float64
 9   IUD                                 858 non-null    float64
 10  IUD (years)                         858 non-null    float64
 11  STDs                                858 non-null    float64
 12  STDs (number)                       858 non-null    float64
 13  STDs:condylomatosis                 858 non-null    float64
 14  STDs:cervical condylomatosis        858 non-null    float64
 15  STDs:vaginal condylomatosis         858 non-null    float64
 16  STDs:vulvo-perineal condylomatosis  858 non-null    float64
 17  STDs:syphilis                       858 non-null    float64
 18  STDs:pelvic inflammatory disease    858 non-null    float64
 19  STDs:genital herpes                 858 non-null    float64
 20  STDs:molluscum contagiosum          858 non-null    float64
 21  STDs:AIDS                           858 non-null    float64
 22  STDs:HIV                            858 non-null    float64
 23  STDs:Hepatitis B                    858 non-null    float64
 24  STDs:HPV                            858 non-null    float64
 25  STDs: Number of diagnosis           858 non-null    int64
 26  STDs: Time since first diagnosis    858 non-null    float64
 27  STDs: Time since last diagnosis     858 non-null    float64
 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: float64(26), int64(10)
memory usage: 241.4 KB
```

The number of NaN Values Per Column

[147]:
```
missing_values = data.isnull().sum() #checking the total number of NaN Values per Column
missing_values
```

[147]:
```
Age                                   0
Number of sexual partners             0
First sexual intercourse              0
Num of pregnancies                    0
Smokes                                0
Smokes (years)                        0
Smokes (packs/year)                   0
Hormonal Contraceptives               0
Hormonal Contraceptives (years)       0
IUD                                   0
IUD (years)                           0
STDs                                  0
STDs (number)                         0
STDs:condylomatosis                   0
STDs:cervical condylomatosis          0
STDs:vaginal condylomatosis           0
STDs:vulvo-perineal condylomatosis    0
STDs:syphilis                         0
STDs:pelvic inflammatory disease      0
STDs:genital herpes                   0
STDs:molluscum contagiosum            0
STDs:AIDS                             0
STDs:HIV                              0
STDs:Hepatitis B                      0
STDs:HPV                              0
STDs: Number of diagnosis             0
STDs: Time since first diagnosis      0
STDs: Time since last diagnosis       0
Dx:Cancer                             0
Dx:CIN                                0
Dx:HPV                                0
Dx                                    0
Hinselmann                            0
Schiller                              0
Citology                              0
```

```
Biopsy                              0
dtype: int64
```

Unique values of each column

[148]:
```
for columns in data:
  print(columns)
  print(data[columns].unique()) #checking and printing all the unqiue values of␣
 ↪each column
  print('-----')
```

```
Age
[18 15 34 52 46 42 51 26 45 44 27 43 40 41 39 37 38 36 35 33 31 32 30 23
 28 29 20 25 21 24 22 48 19 17 16 14 59 79 84 47 13 70 50 49]
-----
Number of sexual partners
[ 4.  1.  5.  3.  2.  6.  7. 15.  8. 10. 28.  9.]
-----
First sexual intercourse
[15. 14. 16. 21. 23. 17. 26. 20. 25. 18. 27. 19. 24. 32. 13. 29. 11. 12.
 22. 28. 10.]
-----
Num of pregnancies
[ 1.  4.  2.  6.  3.  5.  8.  7.  0. 11. 10.]
-----
Smokes
[0. 1.]
-----
Smokes (years)
[ 0.         37.         34.          1.26697291  3.         12.
 18.          7.         19.         21.         15.         13.
 16.          8.          4.         10.         22.         14.
  0.5        11.          9.          2.          5.          6.
  1.         32.         24.         28.         20.          0.16      ]
-----
Smokes (packs/year)
[0.00000000e+00 3.70000000e+01 3.40000000e+00 2.80000000e+00
 4.00000000e-02 5.13202128e-01 2.40000000e+00 6.00000000e+00
 9.00000000e+00 1.60000000e+00 1.90000000e+01 2.10000000e+01
 3.20000000e-01 2.60000000e+00 8.00000000e-01 1.50000000e+01
 2.00000000e+00 5.70000000e+00 1.00000000e+00 3.30000000e+00
 3.50000000e+00 1.20000000e+01 2.50000000e-02 2.75000000e+00
 2.00000000e-01 1.40000000e+00 5.00000000e+00 2.10000000e+00
 7.00000000e-01 1.20000000e+00 7.50000000e+00 1.25000000e+00
 3.00000000e+00 7.50000000e-01 1.00000000e-01 8.00000000e+00
 2.25000000e+00 3.00000000e-03 7.00000000e+00 4.50000000e-01
 1.50000000e-01 5.00000000e-02 2.50000000e-01 4.80000000e+00
 4.50000000e+00 4.00000000e-01 3.70000000e-01 2.20000000e+00
```

```
 1.60000000e-01 9.00000000e-01 2.20000000e+01 1.35000000e+00
 5.00000000e-01 2.50000000e+00 4.00000000e+00 1.30000000e+00
 1.65000000e+00 2.70000000e+00 1.00000000e-03 7.60000000e+00
 5.50000000e+00 3.00000000e-01]
-----
Hormonal Contraceptives
[0. 1.]
-----
Hormonal Contraceptives (years)
[ 0.          3.         15.          2.          8.         10.
  5.          0.25        7.         22.         19.          0.5
  1.          0.58        9.         13.         11.          4.
 12.         16.          0.33        0.16       14.          0.08
  2.28220052  0.66        6.          1.5         0.42        0.67
  0.75        2.5         4.5         6.5         0.17       20.
  3.5         0.41       30.         17.        ]
-----
IUD
[0. 1.]
-----
IUD (years)
[ 0.    7.    5.    8.    6.    1.    0.58  2.   19.    0.5  17.    0.08
  0.25 10.   11.    3.   15.   12.    9.    1.5   0.91  4.    0.33  0.41
  0.16  0.17]
-----
STDs
[0. 1.]
-----
STDs (number)
[0. 2. 1. 3. 4.]
-----
STDs:condylomatosis
[0. 1.]
-----
STDs:cervical condylomatosis
[0.]
-----
STDs:vaginal condylomatosis
[0. 1.]
-----
STDs:vulvo-perineal condylomatosis
[0. 1.]
-----
STDs:syphilis
[0. 1.]
-----
STDs:pelvic inflammatory disease
[0. 1.]
```

```
-----
STDs:genital herpes
[0. 1.]
-----
STDs:molluscum contagiosum
[0. 1.]
-----
STDs:AIDS
[0.]
-----
STDs:HIV
[0. 1.]
-----
STDs:Hepatitis B
[0. 1.]
-----
STDs:HPV
[0. 1.]
-----
STDs: Number of diagnosis
[0 1 3 2]
-----
STDs: Time since first diagnosis
[ 1. 21.  2. 15. 19.  3. 12. 11.  9.  7.  8. 16.  6.  5. 10.  4. 22. 18.]
-----
STDs: Time since last diagnosis
[ 1. 21.  2. 15. 19.  3. 12. 11.  9.  7.  8. 16.  6.  5. 10.  4. 22. 18.]
-----
Dx:Cancer
[0 1]
-----
Dx:CIN
[0 1]
-----
Dx:HPV
[0 1]
-----
Dx
[0 1]
-----
Hinselmann
[0 1]
-----
Schiller
[0 1]
-----
Citology
[0 1]
```

```
-----
Biopsy
[0 1]
-----
```

STDs: Time since first diagnosis and STDs: Time since last diagnosis have to much NaN values, we cannot just imputate its missing data so it will be better to drop those columns

[149]:
```
data = data.drop(['STDs: Time since first diagnosis', 'STDs: Time since last␣
↪diagnosis'], axis=1) #dropping the column
data.columns
```

[149]:
```
Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
       'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:cervical condylomatosis', 'STDs:vaginal condylomatosis',
       'STDs:vulvo-perineal condylomatosis', 'STDs:syphilis',
       'STDs:pelvic inflammatory disease', 'STDs:genital herpes',
       'STDs:molluscum contagiosum', 'STDs:AIDS', 'STDs:HIV',
       'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

The Column STDs: AIDS and STDs:cervical condylomatosis only have NaNs and 0s, we are going to drop it

[150]:
```
data = data.drop(columns = ['STDs:AIDS', 'STDs:cervical condylomatosis'])␣
↪#dropping unecessary columns
data.columns
```

[150]:
```
Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
       'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:vaginal condylomatosis', 'STDs:vulvo-perineal condylomatosis',
       'STDs:syphilis', 'STDs:pelvic inflammatory disease',
       'STDs:genital herpes', 'STDs:molluscum contagiosum', 'STDs:HIV',
       'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

Most of our columns that have missing values are dichotomous variable, which means they only have 2 levels of values which is 'Yes' or 'No', in this case we have '0' or '1', so instead of mean imputation, we are going with mode imputation for the NaN values.

The rest of are columns are in Interval Level of measurement, they cannot have vaues with decimals, only whole numbers that's why mode imputation is also used

[151]:
```
for x in data:
  data[x] = data[x].fillna(data[x].mode()[0]) #mode imputation
```

[152]:
```
nan_values = data.isnull().sum() #checking again for null values
nan_values
```

[152]:
```
Age                                 0
Number of sexual partners           0
First sexual intercourse            0
Num of pregnancies                  0
Smokes                              0
Smokes (years)                      0
Smokes (packs/year)                 0
Hormonal Contraceptives             0
Hormonal Contraceptives (years)     0
IUD                                 0
IUD (years)                         0
STDs                                0
STDs (number)                       0
STDs:condylomatosis                 0
STDs:vaginal condylomatosis         0
STDs:vulvo-perineal condylomatosis  0
STDs:syphilis                       0
STDs:pelvic inflammatory disease    0
STDs:genital herpes                 0
STDs:molluscum contagiosum          0
STDs:HIV                            0
STDs:Hepatitis B                    0
STDs:HPV                            0
STDs: Number of diagnosis           0
Dx:Cancer                           0
Dx:CIN                              0
Dx:HPV                              0
Dx                                  0
Hinselmann                          0
Schiller                            0
Citology                            0
Biopsy                              0
dtype: int64
```

[153]:
```
data
```

[153]:

	Age	Number of sexual partners	First sexual intercourse	\
0	18	4.0	15.0	
1	15	1.0	14.0	
2	34	1.0	15.0	
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	0.00	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 32 columns]
```

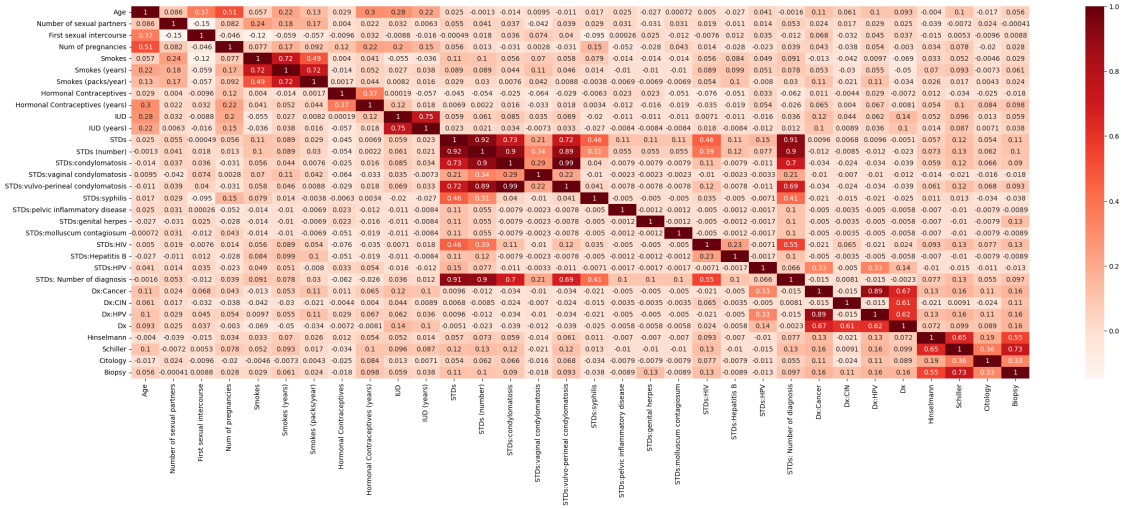
[154]:
```
data.dtypes
```

[154]:
```
Age                                   int64
Number of sexual partners           float64
First sexual intercourse            float64
Num of pregnancies                  float64
Smokes                              float64
Smokes (years)                      float64
Smokes (packs/year)                 float64
Hormonal Contraceptives             float64
Hormonal Contraceptives (years)     float64
IUD                                 float64
IUD (years)                         float64
STDs                                float64
STDs (number)                       float64
STDs:condylomatosis                 float64
STDs:vaginal condylomatosis         float64
STDs:vulvo-perineal condylomatosis  float64
STDs:syphilis                       float64
STDs:pelvic inflammatory disease    float64
STDs:genital herpes                 float64
STDs:molluscum contagiosum          float64
STDs:HIV                            float64
STDs:Hepatitis B                    float64
STDs:HPV                            float64
```

```
STDs: Number of diagnosis          int64
Dx:Cancer                          int64
Dx:CIN                             int64
Dx:HPV                             int64
Dx                                 int64
Hinselmann                         int64
Schiller                           int64
Citology                           int64
Biopsy                             int64
dtype: object
```

Correlation of each column using heatmap

[155]:
```
plt.figure(figsize = (30,10))
ax = sns.heatmap(data.corr(), annot = True, cmap = 'Reds')
```



[272]:
```
data.corr()
```

[272]:

	Age	Number of sexual partners	\
Age	1.000000	0.085971	
Number of sexual partners	0.085971	1.000000	
First sexual intercourse	0.365248	-0.147280	
Num of pregnancies	0.514977	0.082388	
Smokes	0.057204	0.236858	
Smokes (years)	0.218261	0.175729	
Smokes (packs/year)	0.131861	0.174968	
Hormonal Contraceptives	0.029201	0.004027	
Hormonal Contraceptives (years)	0.298892	0.021525	
IUD	0.279429	0.032460	
IUD (years)	0.215427	0.006252	

STDs	0.025241	0.055370
STDs (number)	-0.001330	0.041459
STDs:condylomatosis	-0.013751	0.036925
STDs:vaginal condylomatosis	0.009505	-0.042120
STDs:vulvo-perineal condylomatosis	-0.011499	0.038992
STDs:syphilis	0.017457	0.028646
STDs:pelvic inflammatory disease	0.024854	0.030929
STDs:genital herpes	-0.027433	-0.031413
STDs:molluscum contagiosum	0.000722	0.030929
STDs:HIV	0.005009	0.018752
STDs:Hepatitis B	-0.027433	-0.010633
STDs:HPV	0.040861	0.014360
STDs: Number of diagnosis	-0.001606	0.053056
Dx:Cancer	0.110340	0.023699
Dx:CIN	0.061443	0.016669
Dx:HPV	0.101722	0.028646
Dx	0.092635	0.024597
Hinselmann	-0.003967	-0.039098
Schiller	0.103283	-0.007230
Citology	-0.016862	0.024067
Biopsy	0.055956	-0.000408

	First sexual intercourse	\
Age	0.365248	
Number of sexual partners	-0.147280	
First sexual intercourse	1.000000	
Num of pregnancies	-0.046099	
Smokes	-0.123017	
Smokes (years)	-0.058620	
Smokes (packs/year)	-0.057013	
Hormonal Contraceptives	-0.009563	
Hormonal Contraceptives (years)	0.031976	
IUD	-0.008826	
IUD (years)	-0.015697	
STDs	-0.000494	
STDs (number)	0.017948	
STDs:condylomatosis	0.035761	
STDs:vaginal condylomatosis	0.073945	
STDs:vulvo-perineal condylomatosis	0.039933	
STDs:syphilis	-0.094885	
STDs:pelvic inflammatory disease	0.000256	
STDs:genital herpes	0.024691	
STDs:molluscum contagiosum	-0.011961	
STDs:HIV	-0.007627	
STDs:Hepatitis B	0.012473	
STDs:HPV	0.034938	
STDs: Number of diagnosis	-0.011617	

```
Dx:Cancer                                      0.067996
Dx:CIN                                        -0.031960
Dx:HPV                                         0.044727
Dx                                             0.036664
Hinselmann                                    -0.015311
Schiller                                       0.005275
Citology                                      -0.009594
Biopsy                                         0.008771

                                  Num of pregnancies    Smokes  \
Age                                         0.514977  0.057204
Number of sexual partners                   0.082388  0.236858
First sexual intercourse                   -0.046099 -0.123017
Num of pregnancies                          1.000000  0.077363
Smokes                                      0.077363  1.000000
Smokes (years)                              0.172084  0.723572
Smokes (packs/year)                         0.092214  0.493843
Hormonal Contraceptives                     0.116944  0.004036
Hormonal Contraceptives (years)             0.221456  0.040917
IUD                                         0.198134 -0.055115
IUD (years)                                 0.148692 -0.035798
STDs                                        0.055698  0.111289
STDs (number)                               0.012938  0.100117
STDs:condylomatosis                        -0.031189  0.055674
STDs:vaginal condylomatosis                 0.002754  0.069651
STDs:vulvo-perineal condylomatosis         -0.030813  0.058468
STDs:syphilis                               0.150551  0.079358
STDs:pelvic inflammatory disease           -0.052239 -0.013974
STDs:genital herpes                        -0.028411 -0.013974
STDs:molluscum contagiosum                  0.043074 -0.013974
STDs:HIV                                    0.014401  0.056151
STDs:Hepatitis B                           -0.028411  0.083503
STDs:HPV                                   -0.023343  0.049193
STDs: Number of diagnosis                   0.039195  0.090725
Dx:Cancer                                   0.042765 -0.013470
Dx:CIN                                     -0.037752 -0.042119
Dx:HPV                                      0.054111  0.009737
Dx                                         -0.003034 -0.069396
Hinselmann                                  0.033987  0.033333
Schiller                                    0.077526  0.052028
Citology                                   -0.020131 -0.004639
Biopsy                                      0.027959  0.028724

                                  Smokes (years)  Smokes (packs/year)  \
Age                                     0.218261             0.131861
Number of sexual partners               0.175729             0.174968
First sexual intercourse               -0.058620            -0.057013
```

Num of pregnancies	0.172084	0.092214
Smokes	0.723572	0.493843
Smokes (years)	1.000000	0.724320
Smokes (packs/year)	0.724320	1.000000
Hormonal Contraceptives	-0.013888	0.001713
Hormonal Contraceptives (years)	0.052436	0.043803
IUD	0.027492	0.008226
IUD (years)	0.038061	0.016292
STDs	0.089300	0.029252
STDs (number)	0.088605	0.030247
STDs:condylomatosis	0.043504	0.007599
STDs:vaginal condylomatosis	0.114655	0.041939
STDs:vulvo-perineal condylomatosis	0.045561	0.008761
STDs:syphilis	0.013850	-0.003754
STDs:pelvic inflammatory disease	-0.010111	-0.006901
STDs:genital herpes	-0.010111	-0.006901
STDs:molluscum contagiosum	-0.010111	-0.006901
STDs:HIV	0.088930	0.053995
STDs:Hepatitis B	0.099313	0.101342
STDs:HPV	0.051201	-0.008015
STDs: Number of diagnosis	0.078303	0.029912
Dx:Cancer	0.052859	0.107229
Dx:CIN	-0.030476	-0.020800
Dx:HPV	0.055398	0.109118
Dx	-0.050213	-0.034270
Hinselmann	0.070352	0.026086
Schiller	0.093479	0.017200
Citology	-0.007275	0.004250
Biopsy	0.061204	0.024487

	Hormonal Contraceptives \
Age	0.029201
Number of sexual partners	0.004027
First sexual intercourse	-0.009563
Num of pregnancies	0.116944
Smokes	0.004036
Smokes (years)	-0.013888
Smokes (packs/year)	0.001713
Hormonal Contraceptives	1.000000
Hormonal Contraceptives (years)	0.370696
IUD	0.000188
IUD (years)	-0.056548
STDs	-0.045460
STDs (number)	-0.053642
STDs:condylomatosis	-0.025116
STDs:vaginal condylomatosis	-0.064390
STDs:vulvo-perineal condylomatosis	-0.029001

STDs:syphilis	-0.006252
STDs:pelvic inflammatory disease	0.023085
STDs:genital herpes	0.023085
STDs:molluscum contagiosum	-0.050547
STDs:HIV	-0.076371
STDs:Hepatitis B	-0.050547
STDs:HPV	0.032666
STDs: Number of diagnosis	-0.062199
Dx:Cancer	0.011278
Dx:CIN	-0.004397
Dx:HPV	0.028808
Dx	-0.007245
Hinselmann	0.012360
Schiller	-0.034002
Citology	-0.025116
Biopsy	-0.018015

	Hormonal Contraceptives (years)	IUD	\
Age	0.298892	0.279429	
Number of sexual partners	0.021525	0.032460	
First sexual intercourse	0.031976	-0.008826	
Num of pregnancies	0.221456	0.198134	
Smokes	0.040917	-0.055115	
Smokes (years)	0.052436	0.027492	
Smokes (packs/year)	0.043803	0.008226	
Hormonal Contraceptives	0.370696	0.000188	
Hormonal Contraceptives (years)	1.000000	0.115456	
IUD	0.115456	1.000000	
IUD (years)	0.017955	0.749288	
STDs	0.006918	0.059427	
STDs (number)	0.002236	0.060591	
STDs:condylomatosis	0.016465	0.084794	
STDs:vaginal condylomatosis	-0.032782	0.035484	
STDs:vulvo-perineal condylomatosis	0.018090	0.069399	
STDs:syphilis	0.003385	-0.020393	
STDs:pelvic inflammatory disease	-0.011613	-0.011179	
STDs:genital herpes	-0.016362	-0.011179	
STDs:molluscum contagiosum	-0.018737	-0.011179	
STDs:HIV	-0.035063	0.007118	
STDs:Hepatitis B	-0.018737	-0.011179	
STDs:HPV	0.054142	-0.015819	
STDs: Number of diagnosis	-0.025662	0.035791	
Dx:Cancer	0.064993	0.117166	
Dx:CIN	0.003972	0.043708	
Dx:HPV	0.066509	0.062142	
Dx	-0.008054	0.135778	
Hinselmann	0.054264	0.052108	

```
Schiller                                          0.101250  0.096089
Citology                                          0.084429  0.013292
Biopsy                                            0.097937  0.059231

                                  …  STDs:HPV  STDs: Number of diagnosis  \
Age                               …  0.040861                 -0.001606
Number of sexual partners         …  0.014360                  0.053056
First sexual intercourse          …  0.034938                 -0.011617
Num of pregnancies                … -0.023343                  0.039195
Smokes                            …  0.049193                  0.090725
Smokes (years)                    …  0.051201                  0.078303
Smokes (packs/year)               … -0.008015                  0.029912
Hormonal Contraceptives           …  0.032666                 -0.062199
Hormonal Contraceptives (years)   …  0.054142                 -0.025662
IUD                               … -0.015819                  0.035791
IUD (years)                       … -0.011853                  0.012191
STDs                              …  0.151787                  0.907805
STDs (number)                     …  0.077165                  0.898446
STDs:condylomatosis               … -0.011238                  0.701701
STDs:vaginal condylomatosis       … -0.003308                  0.206557
STDs:vulvo-perineal condylomatosis … -0.011103                 0.693256
STDs:syphilis                     … -0.007076                  0.414913
STDs:pelvic inflammatory disease  … -0.001651                  0.103097
STDs:genital herpes               … -0.001651                  0.103097
STDs:molluscum contagiosum        … -0.001651                  0.103097
STDs:HIV                          … -0.007076                  0.549393
STDs:Hepatitis B                  … -0.001651                  0.103097
STDs:HPV                          …  1.000000                  0.065957
STDs: Number of diagnosis         …  0.065957                  1.000000
Dx:Cancer                         …  0.330203                 -0.015423
Dx:CIN                            … -0.004977                  0.008070
Dx:HPV                            …  0.330203                 -0.015423
Dx                                …  0.138371                 -0.002289
Hinselmann                        … -0.009968                  0.076787
Schiller                          … -0.014850                  0.130873
Citology                          … -0.011238                  0.055114
Biopsy                            … -0.012650                  0.097449

                                  Dx:Cancer    Dx:CIN    Dx:HPV        Dx  \
Age                                0.110340  0.061443  0.101722  0.092635
Number of sexual partners          0.023699  0.016669  0.028646  0.024597
First sexual intercourse           0.067996 -0.031960  0.044727  0.036664
Num of pregnancies                 0.042765 -0.037752  0.054111 -0.003034
Smokes                            -0.013470 -0.042119  0.009737 -0.069396
Smokes (years)                     0.052859 -0.030476  0.055398 -0.050213
Smokes (packs/year)                0.107229 -0.020800  0.109118 -0.034270
Hormonal Contraceptives            0.011278 -0.004397  0.028808 -0.007245
```

Hormonal Contraceptives (years)	0.064993	0.003972	0.066509	-0.008054
IUD	0.117166	0.043708	0.062142	0.135778
IUD (years)	0.103148	0.008887	0.035869	0.100340
STDs	0.009638	0.006779	0.009638	-0.005129
STDs (number)	-0.012141	-0.008539	-0.012141	-0.022972
STDs:condylomatosis	-0.034034	-0.023938	-0.034034	-0.039440
STDs:vaginal condylomatosis	-0.010018	-0.007046	-0.010018	-0.011610
STDs:vulvo-perineal condylomatosis	-0.033624	-0.023650	-0.033624	-0.038965
STDs:syphilis	-0.021429	-0.015072	-0.021429	-0.024832
STDs:pelvic inflammatory disease	-0.005000	-0.003517	-0.005000	-0.005795
STDs:genital herpes	-0.005000	-0.003517	-0.005000	-0.005795
STDs:molluscum contagiosum	-0.005000	-0.003517	-0.005000	-0.005795
STDs:HIV	-0.021429	0.064753	-0.021429	0.024488
STDs:Hepatitis B	-0.005000	-0.003517	-0.005000	-0.005795
STDs:HPV	0.330203	-0.004977	0.330203	0.138371
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

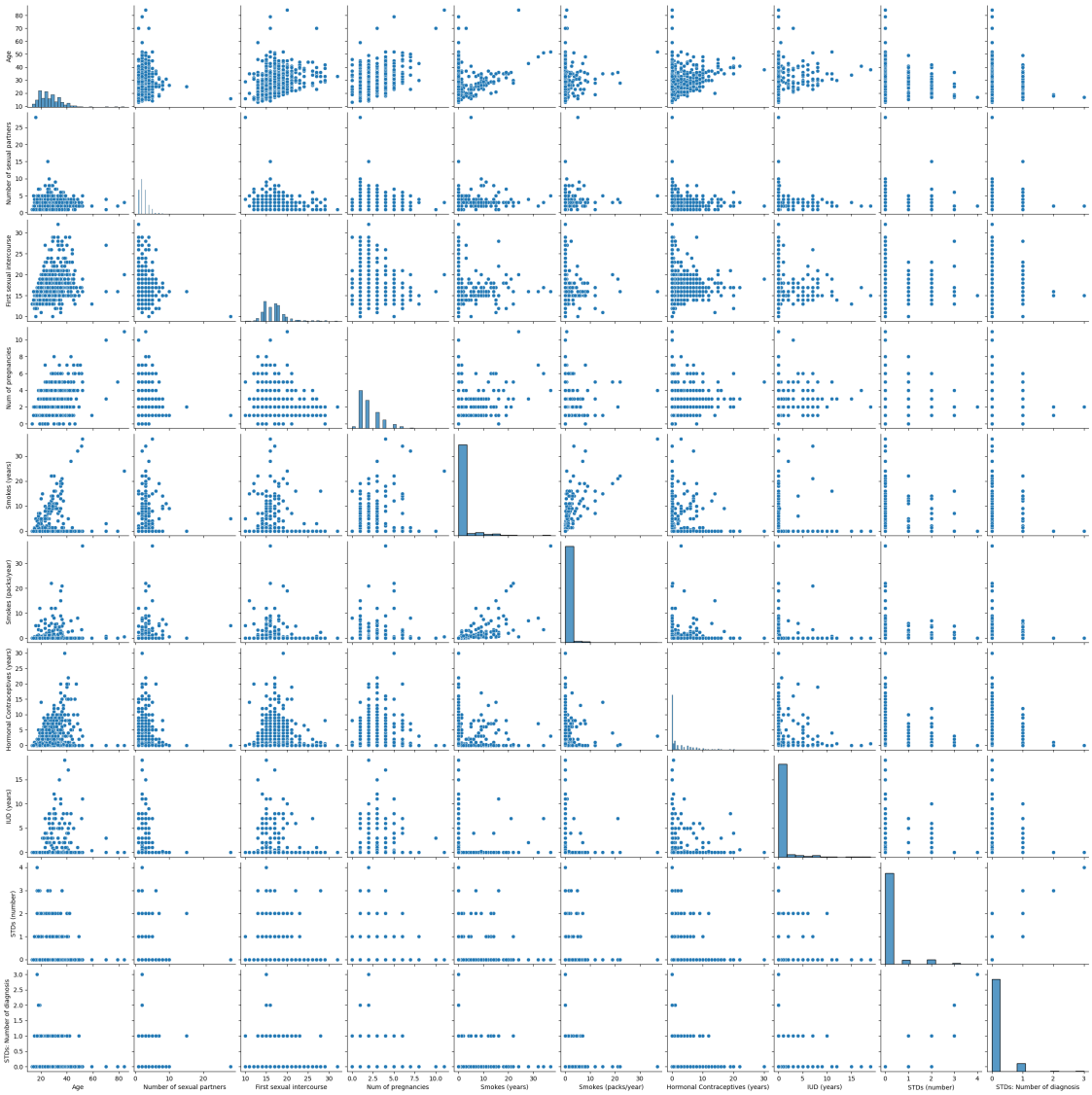
	Hinselmann	Schiller	Citology	Biopsy
Age	-0.003967	0.103283	-0.016862	0.055956
Number of sexual partners	-0.039098	-0.007230	0.024067	-0.000408
First sexual intercourse	-0.015311	0.005275	-0.009594	0.008771
Num of pregnancies	0.033987	0.077526	-0.020131	0.027959
Smokes	0.033333	0.052028	-0.004639	0.028724
Smokes (years)	0.070352	0.093479	-0.007275	0.061204
Smokes (packs/year)	0.026086	0.017200	0.004250	0.024487
Hormonal Contraceptives	0.012360	-0.034002	-0.025116	-0.018015
Hormonal Contraceptives (years)	0.054264	0.101250	0.084429	0.097937
IUD	0.052108	0.096089	0.013292	0.059231
IUD (years)	0.014132	0.087032	0.007103	0.038176
STDs	0.056599	0.117552	0.053889	0.114148
STDs (number)	0.073186	0.129649	0.061689	0.103153
STDs:condylomatosis	0.058905	0.116795	0.065725	0.090164
STDs:vaginal condylomatosis	-0.014114	-0.021026	-0.015912	-0.017911
STDs:vulvo-perineal condylomatosis	0.060651	0.119714	0.067686	0.092548
STDs:syphilis	0.010925	0.012965	-0.034034	-0.038311
STDs:pelvic inflammatory disease	-0.007044	-0.010495	-0.007942	-0.008940
STDs:genital herpes	-0.007044	-0.010495	-0.007942	0.130523
STDs:molluscum contagiosum	-0.007044	-0.010495	-0.007942	-0.008940
STDs:HIV	0.093151	0.128842	0.076576	0.127702

```
STDs:Hepatitis B                    -0.007044 -0.010495 -0.007942 -0.008940
STDs:HPV                            -0.009968 -0.014850 -0.011238 -0.012650
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

[32 rows x 32 columns]
```

[112]:
```
sns.pairplot(data)
```

[112]:
```
<seaborn.axisgrid.PairGrid at 0x7dca564bd5d0>
```



According to our dataset, we can use the 'Hinselmann', 'Schiller', 'Biopsy', and 'Citology' as target variable, so we can use it for our y in our model, for our x, and since our Dataset focuses on cancer, we will have Dx:Cancer for our x variable

[262]:
```
X_1 = data.drop('Hinselmann', axis=1)
y_1 = data['Hinselmann']
```

[263]:
```
X_train, X_test, y_train, y_test = train_test_split(X_1, y_1, test_size=0.3,␣
↪random_state=102)
```

[264]:
```
model = LogisticRegression()

model.fit(X_train, y_train)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

[264]:
```
LogisticRegression()
```

CLASSIFICATION REPORT

[265]:
```
predictions = model.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
print(accuracy)
print(classification_report(y_test, predictions, target_names = ['with cancer',␣
 ↪'without cancer']))
```

```
0.9651162790697675
                precision    recall  f1-score   support

   with cancer       0.98      0.98      0.98       250
without cancer       0.44      0.50      0.47         8

      accuracy                           0.97       258
     macro avg       0.71      0.74      0.73       258
  weighted avg       0.97      0.97      0.97       258
```

PREDICTIONS

[266]:
```
predictions
```

[266]:
```
array([0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

[271]:
```
sample_number = 0

for x in predictions:
  sample_number +=1
  if x == 1:
    print('According to Model predictions, Sample No.', sample_number, 'will␣
 ↪have cancer by using the Hinselmann Instrument')
```

```
According to Model predictions, Sample No. 5 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 33 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 62 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 71 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 84 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 131 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 194 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 213 will have cancer by using the
Hinselmann Instrument
According to Model predictions, Sample No. 216 will have cancer by using the
Hinselmann Instrument
```

Conclusion

- In this activity, we tried to create a classification using Logistic regression to our given dataset, our dataset contains different habits of the population including some of their personal information. I tried creating a model to predict whether our sample will have a cancer or not using a certain equipment instrument, in this case, I used the Hinselmann as my test variable.

[277]:
```
!apt-get install texlive texlive-xetex texlive-latex-extra pandoc
```

```
Reading package lists… Done
Building dependency tree… Done
Reading state information… Done
pandoc is already the newest version (2.9.2.1-3ubuntu2).
texlive is already the newest version (2021.20220204-1).
texlive-latex-extra is already the newest version (2021.20220204-1).
texlive-xetex is already the newest version (2021.20220204-1).
0 upgraded, 0 newly installed, 0 to remove and 45 not upgraded.
```

[ ]:
```

```