De_Guzman_Hands_on_Activity_11_2_Classification_using_Logistic_Reg

April 28, 2024

1 Hands-on Activity 11.2 Classification using Logistic Regression

1.1 Objective(s):

- This activity aims to demonstrate how to apply simple linear regression analysis to solve regression problem

1.2 Intended Learning Outcomes (ILOs):

- Demonstrate how to solve classification problems using Logistic Regression
- Use the logistic regression model to perform classification

1.3 Resources:

- Jupyter Notebook

1.4 Dataset:

- https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

1.5 Submission Requirements:

- PDF containing initial EDA and Data Wrangling
- PDF showing demonstration of simple linear regression.
- Submit a link to the colab file through the comment section.

1.6 Procedure:

1.6.1 Setup

[365]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

%matplotlib inline
```

[366]:
```
import warnings
warnings.filterwarnings('ignore')
```

[367]:
```
!pip install ucimlrepo
```

```
Requirement already satisfied: ucimlrepo in /usr/local/lib/python3.10/dist-
packages (0.0.6)
```

[368]:
```
from ucimlrepo import fetch_ucirepo

# fetch dataset
cervical_cancer_risk_factors = fetch_ucirepo(id=383)

# data (as pandas dataframes)
ccrf_df = cervical_cancer_risk_factors.data.features

# 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
```

1.6.2 Data Wrangling and Cleaning

[369]:
```
# Checking the shape of the dataset
ccrf_df.shape
```

[369]: (858, 36)

[370]:
```
# Checking general information
ccrf_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           832 non-null    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  STDs: Time since first diagnosis    71 non-null     float64
 27  STDs: Time since last diagnosis     71 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
```

[371]:
```
# Dropping unusable columns
ccrf_df.drop(columns=["STDs: Time since first diagnosis","STDs: Time since last␣
 ↪diagnosis"], inplace=True)
```

[372]:
```
# Dropping Age records with only 1 row in the dataset
counts = ccrf_df['Age'].value_counts().sort_values()

for i in list(counts.index):
  if counts[i] < 2:
    ccrf_df.drop(ccrf_df[ccrf_df['Age'] == i].index, axis=0, inplace=True)
```

[373]:
```
# Listing all NaN columns and filtering out the boolean-like datatypes
nan_cols = list(ccrf_df[ccrf_df.columns[ccrf_df.isna().any()]].columns)

for i in nan_cols:
  if len(list(ccrf_df[i].unique())) <= 3:
    print(i)
    print(list(ccrf_df[i].unique()))
```

```
Smokes
[0.0, 1.0, nan]
Hormonal Contraceptives
[0.0, 1.0, nan]
IUD
[0.0, 1.0, nan]
STDs
[0.0, 1.0, nan]
STDs:condylomatosis
[0.0, 1.0, nan]
STDs:cervical condylomatosis
[0.0, nan]
STDs:vaginal condylomatosis
[0.0, nan, 1.0]
STDs:vulvo-perineal condylomatosis
[0.0, 1.0, nan]
STDs:syphilis
[0.0, 1.0, nan]
STDs:pelvic inflammatory disease
[0.0, nan, 1.0]
STDs:genital herpes
[0.0, nan, 1.0]
STDs:molluscum contagiosum
[0.0, nan, 1.0]
STDs:AIDS
[0.0, nan]
STDs:HIV
[0.0, 1.0, nan]
```

```
STDs:Hepatitis B
[0.0, nan, 1.0]
STDs:HPV
[0.0, nan, 1.0]
```

[374]:
```
# Dropping columns with only 0 and NaN values
ccrf_df.drop(['STDs:cervical condylomatosis', 'STDs:AIDS'], axis=1,␣
 ↪inplace=True)
```

[375]:
```
# Relisting all NaN columns and separating the boolean-like datatypes
nan_cols = list(ccrf_df[ccrf_df.columns[ccrf_df.isna().any()]].columns)

nan_mean = []

for i in nan_cols:
  if len(list(ccrf_df[i].unique())) <= 3:
    print(i)
    print(list(ccrf_df[i].unique()))
  else:
    nan_mean.append(i)
```

```
Smokes
[0.0, 1.0, nan]
Hormonal Contraceptives
[0.0, 1.0, nan]
IUD
[0.0, 1.0, nan]
STDs
[0.0, 1.0, nan]
STDs:condylomatosis
[0.0, 1.0, nan]
STDs:vaginal condylomatosis
[0.0, nan, 1.0]
STDs:vulvo-perineal condylomatosis
[0.0, 1.0, nan]
STDs:syphilis
[0.0, 1.0, nan]
STDs:pelvic inflammatory disease
[0.0, nan, 1.0]
STDs:genital herpes
[0.0, nan, 1.0]
STDs:molluscum contagiosum
[0.0, nan, 1.0]
STDs:HIV
[0.0, 1.0, nan]
STDs:Hepatitis B
[0.0, nan, 1.0]
STDs:HPV
```

```
[0.0, nan, 1.0]
```

[376]:
```
# Checking the frequency of records for each non-boolean column
for i in nan_mean:
  print(ccrf_df[i].value_counts())
```

```
Number of sexual partners
2.0     268
3.0     206
1.0     205
4.0      78
5.0      44
6.0       9
7.0       7
8.0       4
15.0      1
10.0      1
28.0      1
9.0       1
Name: count, dtype: int64
First sexual intercourse
15.0    163
17.0    148
18.0    137
16.0    120
14.0     79
19.0     60
20.0     36
13.0     23
21.0     20
23.0      9
22.0      9
26.0      7
12.0      6
27.0      6
24.0      6
29.0      5
28.0      3
11.0      2
25.0      2
10.0      2
32.0      1
Name: count, dtype: int64
Num of pregnancies
1.0     270
2.0     240
3.0     138
4.0      74
```

```
5.0      34
6.0      17
0.0      15
7.0       5
8.0       2
10.0      1
Name: count, dtype: int64
Smokes (years)
0.000000     717
1.266973      15
5.000000       9
9.000000       9
1.000000       8
3.000000       7
2.000000       7
8.000000       6
7.000000       6
16.000000      6
11.000000      5
4.000000       5
10.000000      5
14.000000      4
15.000000      4
6.000000       4
13.000000      3
0.500000       3
19.000000      3
12.000000      3
22.000000      2
37.000000      1
21.000000      1
18.000000      1
32.000000      1
28.000000      1
20.000000      1
0.160000       1
Name: count, dtype: int64
Smokes (packs/year)
0.000000     717
0.513202      17
1.000000       6
3.000000       5
2.000000       4
              …
37.000000      1
2.250000       1
0.003000       1
0.450000       1
```

```
0.300000       1
Name: count, Length: 61, dtype: int64
Hormonal Contraceptives (years)
0.000000     264
1.000000      77
0.250000      41
2.000000      40
3.000000      39
5.000000      33
0.080000      25
0.500000      25
6.000000      24
4.000000      22
7.000000      21
8.000000      18
0.160000      16
9.000000      12
10.000000     11
0.330000       9
0.420000       8
0.750000       7
15.000000      6
0.580000       6
0.660000       6
12.000000      4
20.000000      3
1.500000       3
0.670000       2
13.000000      2
11.000000      2
2.282201       2
14.000000      2
19.000000      2
16.000000      2
22.000000      1
2.500000       1
4.500000       1
6.500000       1
0.170000       1
3.500000       1
0.410000       1
30.000000      1
17.000000      1
Name: count, dtype: int64
IUD (years)
0.00     653
3.00      11
2.00      10
```

```
5.00       9
1.00       8
8.00       7
7.00       6
6.00       5
4.00       5
11.00      3
0.50       2
0.08       2
0.91       1
0.33       1
9.00       1
0.16       1
1.50       1
0.25       1
12.00      1
15.00      1
10.00      1
17.00      1
19.00      1
0.58       1
0.17       1
Name: count, dtype: int64
STDs (number)
0.0    667
2.0     37
1.0     34
3.0      7
4.0      1
Name: count, dtype: int64
```

[377]:
```
# Since the distribution is more spread out between the high frequency records
# We will be using mean to fill the missing values
mean_cols = ['Number of sexual partners', 'First sexual intercourse', 'Num of␣
 ↪pregnancies']

for i in mean_cols:
  nan_mean.remove(i)

for i in mean_cols:
  ave = ccrf_df[ccrf_df[i].isnull()]['Age'].apply(
      lambda x: ccrf_df[ccrf_df['Age'] == x][i].mean()
  )
  ccrf_df[i].fillna(ave, inplace=True)
  ccrf_df[i] = ccrf_df[i].astype('int64')
  nan_cols.remove(i)
```

[378]:
```
# For the boolean-like values, we will be using mode instead
for i in nan_mean:
  mod = ccrf_df[ccrf_df[i].isnull()]['Age'].apply(
      lambda x: ccrf_df[ccrf_df['Age'] == x][i].mode()
  )
  ccrf_df[i].fillna(mod[0], inplace=True)
```

[379]:
```
for i in nan_mean:
  nan_cols.remove(i)

for i in nan_cols:
  mod = ccrf_df[ccrf_df[i].isnull()]['Age'].apply(
      lambda x: ccrf_df[ccrf_df['Age'] == x][i].mode()
  )
  ccrf_df[i].fillna(mod[0], inplace=True)
```

[380]:
```
# Rechecking the DataFrame
ccrf_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 851 entries, 0 to 857
Data columns (total 32 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 851 non-null    int64
 1   Number of sexual partners           851 non-null    int64
 2   First sexual intercourse            851 non-null    int64
 3   Num of pregnancies                  851 non-null    int64
 4   Smokes                              851 non-null    float64
 5   Smokes (years)                      851 non-null    float64
 6   Smokes (packs/year)                 851 non-null    float64
 7   Hormonal Contraceptives             851 non-null    float64
 8   Hormonal Contraceptives (years)     851 non-null    float64
 9   IUD                                 851 non-null    float64
 10  IUD (years)                         851 non-null    float64
 11  STDs                                851 non-null    float64
 12  STDs (number)                       851 non-null    float64
 13  STDs:condylomatosis                 851 non-null    float64
 14  STDs:vaginal condylomatosis         851 non-null    float64
 15  STDs:vulvo-perineal condylomatosis  851 non-null    float64
 16  STDs:syphilis                       851 non-null    float64
 17  STDs:pelvic inflammatory disease    851 non-null    float64
 18  STDs:genital herpes                 851 non-null    float64
 19  STDs:molluscum contagiosum          851 non-null    float64
 20  STDs:HIV                            851 non-null    float64
 21  STDs:Hepatitis B                    851 non-null    float64
 22  STDs:HPV                            851 non-null    float64
```

```
 23  STDs: Number of diagnosis           851 non-null    int64
 24  Dx:Cancer                           851 non-null    int64
 25  Dx:CIN                              851 non-null    int64
 26  Dx:HPV                              851 non-null    int64
 27  Dx                                  851 non-null    int64
 28  Hinselmann                          851 non-null    int64
 29  Schiller                            851 non-null    int64
 30  Citology                            851 non-null    int64
 31  Biopsy                              851 non-null    int64
dtypes: float64(19), int64(13)
memory usage: 219.4 KB
```

[381]:
```
# Converting the boolean-like columns to categorical
for i in ccrf_df.columns:
  if list(ccrf_df[i].unique()) == [0.0, 1.0]:
    ccrf_df[i] = ccrf_df[i].astype('category')

ccrf_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 851 entries, 0 to 857
Data columns (total 32 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 851 non-null    int64
 1   Number of sexual partners           851 non-null    int64
 2   First sexual intercourse            851 non-null    int64
 3   Num of pregnancies                  851 non-null    int64
 4   Smokes                              851 non-null    category
 5   Smokes (years)                      851 non-null    float64
 6   Smokes (packs/year)                 851 non-null    float64
 7   Hormonal Contraceptives             851 non-null    category
 8   Hormonal Contraceptives (years)     851 non-null    float64
 9   IUD                                 851 non-null    category
 10  IUD (years)                         851 non-null    float64
 11  STDs                                851 non-null    category
 12  STDs (number)                       851 non-null    float64
 13  STDs:condylomatosis                 851 non-null    category
 14  STDs:vaginal condylomatosis         851 non-null    category
 15  STDs:vulvo-perineal condylomatosis  851 non-null    category
 16  STDs:syphilis                       851 non-null    category
 17  STDs:pelvic inflammatory disease    851 non-null    category
 18  STDs:genital herpes                 851 non-null    category
 19  STDs:molluscum contagiosum          851 non-null    category
 20  STDs:HIV                            851 non-null    category
 21  STDs:Hepatitis B                    851 non-null    category
 22  STDs:HPV                            851 non-null    category
 23  STDs: Number of diagnosis           851 non-null    int64
```

```
 24  Dx:Cancer                           851 non-null    category
 25  Dx:CIN                              851 non-null    category
 26  Dx:HPV                              851 non-null    category
 27  Dx                                  851 non-null    category
 28  Hinselmann                          851 non-null    category
 29  Schiller                            851 non-null    category
 30  Citology                            851 non-null    category
 31  Biopsy                              851 non-null    category
dtypes: category(22), float64(5), int64(5)
memory usage: 94.1 KB
```

1.6.3 Exploratory Data Analysis

[382]:
```
# Creating a list of columns that are categorical
cols = ccrf_df.columns

num_cols = ccrf_df._get_numeric_data().columns

cat_cols = list(set(cols)-set(num_cols))

cat_cols
```

[382]:
```
['STDs:HIV',
 'Dx:HPV',
 'STDs:genital herpes',
 'Dx:Cancer',
 'STDs:HPV',
 'STDs:syphilis',
 'Biopsy',
 'Dx',
 'IUD',
 'STDs:condylomatosis',
 'STDs',
 'Smokes',
 'Hormonal Contraceptives',
 'STDs:molluscum contagiosum',
 'Dx:CIN',
 'Citology',
 'Schiller',
 'STDs:vaginal condylomatosis',
 'STDs:Hepatitis B',
 'STDs:vulvo-perineal condylomatosis',
 'Hinselmann',
 'STDs:pelvic inflammatory disease']
```

[383]:
```
ccrf_df.corr()
```

[383]:

	Age	Number of sexual partners	\
Age	1.000000	0.096134	
Number of sexual partners	0.096134	1.000000	
First sexual intercourse	0.398508	-0.148443	
Num of pregnancies	0.517323	0.087251	
Smokes	0.039394	0.234972	
Smokes (years)	0.173850	0.178917	
Smokes (packs/year)	0.140875	0.172938	
Hormonal Contraceptives	0.124810	0.018728	
Hormonal Contraceptives (years)	0.329572	0.025183	
IUD	0.294313	0.026920	
IUD (years)	0.226458	0.002687	
STDs	0.036554	0.052349	
STDs (number)	0.007109	0.038665	
STDs:condylomatosis	-0.008067	0.034708	
STDs:vaginal condylomatosis	0.012258	-0.042824	
STDs:vulvo-perineal condylomatosis	-0.005716	0.036804	
STDs:syphilis	0.023119	0.027253	
STDs:pelvic inflammatory disease	0.027830	0.030626	
STDs:genital herpes	-0.028621	-0.031774	
STDs:molluscum contagiosum	0.001776	0.030626	
STDs:HIV	0.009678	0.017348	
STDs:Hepatitis B	-0.028621	-0.010974	
STDs:HPV	0.045525	0.013904	
STDs: Number of diagnosis	0.006700	0.050313	
Dx:Cancer	0.123406	0.022300	
Dx:CIN	0.023543	0.020522	
Dx:HPV	0.114101	0.027253	
Dx	0.077433	0.025772	
Hinselmann	-0.016784	-0.043440	
Schiller	0.067803	-0.011495	
Citology	-0.025413	0.023791	
Biopsy	0.041659	-0.002841	

	First sexual intercourse	\
Age	0.398508	
Number of sexual partners	-0.148443	
First sexual intercourse	1.000000	
Num of pregnancies	-0.061813	
Smokes	-0.129191	
Smokes (years)	-0.070663	
Smokes (packs/year)	-0.057250	
Hormonal Contraceptives	0.027286	
Hormonal Contraceptives (years)	0.032884	
IUD	0.003976	
IUD (years)	-0.017582	
STDs	-0.003167	

STDs (number)	0.015584
STDs:condylomatosis	0.033990
STDs:vaginal condylomatosis	0.073726
STDs:vulvo-perineal condylomatosis	0.038207
STDs:syphilis	-0.096574
STDs:pelvic inflammatory disease	-0.000029
STDs:genital herpes	0.024522
STDs:molluscum contagiosum	-0.012304
STDs:HIV	-0.008892
STDs:Hepatitis B	0.012247
STDs:HPV	0.034700
STDs: Number of diagnosis	-0.014101
Dx:Cancer	0.067100
Dx:CIN	-0.017514
Dx:HPV	0.043718
Dx	0.046541
Hinselmann	-0.017350
Schiller	-0.001775
Citology	-0.011714
Biopsy	0.006744

	Num of pregnancies	Smokes	\
Age	0.517323	0.039394	
Number of sexual partners	0.087251	0.234972	
First sexual intercourse	-0.061813	-0.129191	
Num of pregnancies	1.000000	0.064492	
Smokes	0.064492	1.000000	
Smokes (years)	0.131540	0.731772	
Smokes (packs/year)	0.107429	0.493729	
Hormonal Contraceptives	0.174833	0.012862	
Hormonal Contraceptives (years)	0.237313	0.049817	
IUD	0.218666	-0.065794	
IUD (years)	0.150820	-0.046597	
STDs	0.063663	0.113248	
STDs (number)	0.017135	0.101894	
STDs:condylomatosis	-0.029894	0.056889	
STDs:vaginal condylomatosis	0.001458	0.070408	
STDs:vulvo-perineal condylomatosis	-0.029356	0.059695	
STDs:syphilis	0.155067	0.080456	
STDs:pelvic inflammatory disease	-0.055582	-0.013964	
STDs:genital herpes	-0.030555	-0.013964	
STDs:molluscum contagiosum	0.044525	-0.013964	
STDs:HIV	0.023975	0.057072	
STDs:Hepatitis B	-0.030555	0.084248	
STDs:HPV	-0.025530	0.049727	
STDs: Number of diagnosis	0.046418	0.092415	
Dx:Cancer	0.041851	-0.013080	

Dx:CIN	0.019837	-0.039661
Dx:HPV	0.053769	0.010304
Dx	0.031295	-0.067855
Hinselmann	0.037163	0.020027
Schiller	0.066088	0.035344
Citology	-0.021527	-0.001751
Biopsy	0.032965	0.020386

	Smokes (years)	Smokes (packs/year)	\
Age	0.173850	0.140875	
Number of sexual partners	0.178917	0.172938	
First sexual intercourse	-0.070663	-0.057250	
Num of pregnancies	0.131540	0.107429	
Smokes	0.731772	0.493729	
Smokes (years)	1.000000	0.756261	
Smokes (packs/year)	0.756261	1.000000	
Hormonal Contraceptives	0.022803	0.013656	
Hormonal Contraceptives (years)	0.076611	0.050545	
IUD	-0.000894	0.002299	
IUD (years)	0.005594	0.010524	
STDs	0.099750	0.029432	
STDs (number)	0.098615	0.030416	
STDs:condylomatosis	0.049766	0.007707	
STDs:vaginal condylomatosis	0.122828	0.042018	
STDs:vulvo-perineal condylomatosis	0.051908	0.008870	
STDs:syphilis	0.016946	-0.003697	
STDs:pelvic inflammatory disease	-0.010219	-0.006895	
STDs:genital herpes	-0.010219	-0.006895	
STDs:molluscum contagiosum	-0.010219	-0.006895	
STDs:HIV	0.096700	0.054125	
STDs:Hepatitis B	0.106008	0.101475	
STDs:HPV	0.055122	-0.008004	
STDs: Number of diagnosis	0.087610	0.030078	
Dx:Cancer	0.058383	0.107425	
Dx:CIN	-0.029023	-0.019582	
Dx:HPV	0.061080	0.109316	
Dx	-0.049654	-0.033502	
Hinselmann	0.027587	0.018572	
Schiller	0.045542	0.012742	
Citology	-0.002630	0.005480	
Biopsy	0.030026	0.019554	

	Hormonal Contraceptives	\
Age	0.124810	
Number of sexual partners	0.018728	
First sexual intercourse	0.027286	
Num of pregnancies	0.174833	

Smokes	0.012862
Smokes (years)	0.022803
Smokes (packs/year)	0.013656
Hormonal Contraceptives	1.000000
Hormonal Contraceptives (years)	0.396248
IUD	0.033442
IUD (years)	-0.036943
STDs	-0.028755
STDs (number)	-0.038108
STDs:condylomatosis	-0.013032
STDs:vaginal condylomatosis	-0.059996
STDs:vulvo-perineal condylomatosis	-0.017002
STDs:syphilis	0.001218
STDs:pelvic inflammatory disease	0.024468
STDs:genital herpes	0.024468
STDs:molluscum contagiosum	-0.048083
STDs:HIV	-0.067878
STDs:Hepatitis B	-0.048083
STDs:HPV	0.034623
STDs: Number of diagnosis	-0.046731
Dx:Cancer	0.018492
Dx:CIN	0.017978
Dx:HPV	0.035766
Dx	0.011599
Hinselmann	0.031303
Schiller	-0.009485
Citology	-0.039697
Biopsy	-0.001293

	Hormonal Contraceptives (years)	IUD	\
Age	0.329572	0.294313	
Number of sexual partners	0.025183	0.026920	
First sexual intercourse	0.032884	0.003976	
Num of pregnancies	0.237313	0.218666	
Smokes	0.049817	-0.065794	
Smokes (years)	0.076611	-0.000894	
Smokes (packs/year)	0.050545	0.002299	
Hormonal Contraceptives	0.396248	0.033442	
Hormonal Contraceptives (years)	1.000000	0.164459	
IUD	0.164459	1.000000	
IUD (years)	0.019166	0.740203	
STDs	0.004775	0.058620	
STDs (number)	0.000246	0.059850	
STDs:condylomatosis	0.014953	0.084224	
STDs:vaginal condylomatosis	-0.033394	0.035316	
STDs:vulvo-perineal condylomatosis	0.016604	0.068826	
STDs:syphilis	0.002401	-0.020799	

```
STDs:pelvic inflammatory disease                    -0.011897 -0.011276
STDs:genital herpes                                 -0.016667 -0.011276
STDs:molluscum contagiosum                          -0.019053 -0.011276
STDs:HIV                                            -0.036219  0.006728
STDs:Hepatitis B                                    -0.019053 -0.011276
STDs:HPV                                             0.054049 -0.015956
STDs: Number of diagnosis                           -0.027758  0.035041
Dx:Cancer                                            0.064284  0.116834
Dx:CIN                                               0.010248  0.009019
Dx:HPV                                               0.083980  0.061781
Dx                                                   0.010709  0.116183
Hinselmann                                           0.057177  0.034055
Schiller                                             0.084960  0.087015
Citology                                             0.057271  0.032660
Biopsy                                               0.076637  0.046396
```

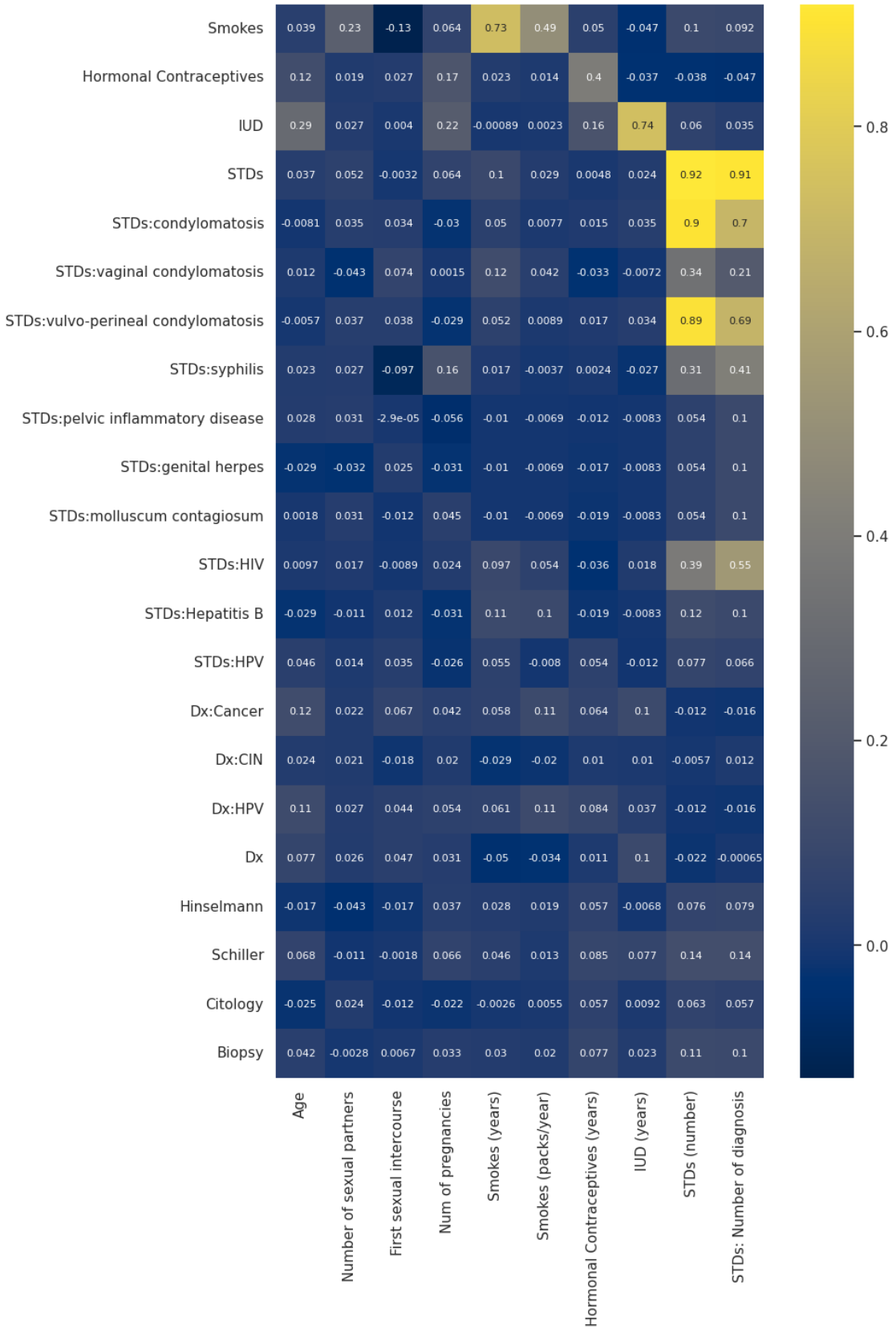
```
                                  …  STDs:HPV  STDs: Number of diagnosis  \
Age                               …  0.045525                   0.006700
Number of sexual partners         …  0.013904                   0.050313
First sexual intercourse          …  0.034700                  -0.014101
Num of pregnancies                … -0.025530                   0.046418
Smokes                            …  0.049727                   0.092415
Smokes (years)                    …  0.055122                   0.087610
Smokes (packs/year)               … -0.008004                   0.030078
Hormonal Contraceptives           …  0.034623                  -0.046731
Hormonal Contraceptives (years)   …  0.054049                  -0.027758
IUD                               … -0.015956                   0.035041
IUD (years)                       … -0.011811                   0.013104
STDs                              …  0.151725                   0.907739
STDs (number)                     …  0.077076                   0.898375
STDs:condylomatosis               … -0.011333                   0.701545
STDs:vaginal condylomatosis       … -0.003335                   0.206469
STDs:vulvo-perineal condylomatosis … -0.011197                  0.693098
STDs:syphilis                     … -0.007135                   0.414744
STDs:pelvic inflammatory disease  … -0.001665                   0.103052
STDs:genital herpes               … -0.001665                   0.103052
STDs:molluscum contagiosum        … -0.001665                   0.103052
STDs:HIV                          … -0.007135                   0.549282
STDs:Hepatitis B                  … -0.001665                   0.103052
STDs:HPV                          …  1.000000                   0.065865
STDs: Number of diagnosis         …  0.065865                   1.000000
Dx:Cancer                         …  0.330177                  -0.015778
Dx:CIN                            … -0.004728                   0.011834
Dx:HPV                            …  0.330177                  -0.015778
Dx                                …  0.141562                  -0.000645
Hinselmann                        … -0.009901                   0.079145
Schiller                          … -0.014643                   0.136406
```

```
Citology                                … -0.011197                    0.056747
Biopsy                                  … -0.012508                    0.101398

                                        Dx:Cancer    Dx:CIN    Dx:HPV        Dx  \
Age                                      0.123406  0.023543  0.114101  0.077433
Number of sexual partners                0.022300  0.020522  0.027253  0.025772
First sexual intercourse                 0.067100 -0.017514  0.043718  0.046541
Num of pregnancies                       0.041851  0.019837  0.053769  0.031295
Smokes                                  -0.013080 -0.039661  0.010304 -0.067855
Smokes (years)                           0.058383 -0.029023  0.061080 -0.049654
Smokes (packs/year)                      0.107425 -0.019582  0.109316 -0.033502
Hormonal Contraceptives                  0.018492  0.017978  0.035766  0.011599
Hormonal Contraceptives (years)          0.064284  0.010248  0.083980  0.010709
IUD                                      0.116834  0.009019  0.061781  0.116183
IUD (years)                              0.104382  0.010001  0.036566  0.103862
STDs                                     0.009259  0.010798  0.009259 -0.003374
STDs (number)                           -0.012499 -0.005737 -0.012499 -0.021749
STDs:condylomatosis                     -0.034324 -0.022747 -0.034324 -0.038917
STDs:vaginal condylomatosis             -0.010102 -0.006695 -0.010102 -0.011453
STDs:vulvo-perineal condylomatosis      -0.033911 -0.022473 -0.033911 -0.038448
STDs:syphilis                           -0.021609 -0.014320 -0.021609 -0.024500
STDs:pelvic inflammatory disease        -0.005042 -0.003341 -0.005042 -0.005717
STDs:genital herpes                     -0.005042 -0.003341 -0.005042 -0.005717
STDs:molluscum contagiosum              -0.005042 -0.003341 -0.005042 -0.005717
STDs:HIV                                -0.021609  0.070308 -0.021609  0.025861
STDs:Hepatitis B                        -0.005042 -0.003341 -0.005042 -0.005717
STDs:HPV                                 0.330177 -0.004728  0.330177  0.141562
STDs: Number of diagnosis               -0.015778  0.011834 -0.015778 -0.000645
Dx:Cancer                                1.000000 -0.014320  0.886488  0.680550
Dx:CIN                                  -0.014320  1.000000 -0.014320  0.584497
Dx:HPV                                   0.886488 -0.014320  1.000000  0.630189
Dx                                       0.680550  0.584497  0.630189  1.000000
Hinselmann                               0.136806 -0.019873  0.136806  0.077000
Schiller                                 0.162374  0.014644  0.162374  0.106942
Citology                                 0.115228 -0.022473  0.115228  0.093885
Biopsy                                   0.164877  0.126060  0.164877  0.166946

                                        Hinselmann  Schiller  Citology    Biopsy
Age                                      -0.016784  0.067803 -0.025413  0.041659
Number of sexual partners                -0.043440 -0.011495  0.023791 -0.002841
First sexual intercourse                 -0.017350 -0.001775 -0.011714  0.006744
Num of pregnancies                        0.037163  0.066088 -0.021527  0.032965
Smokes                                    0.020027  0.035344 -0.001751  0.020386
Smokes (years)                            0.027587  0.045542 -0.002630  0.030026
Smokes (packs/year)                       0.018572  0.012742  0.005480  0.019554
Hormonal Contraceptives                   0.031303 -0.009485 -0.039697 -0.001293
Hormonal Contraceptives (years)           0.057177  0.084960  0.057271  0.076637
```

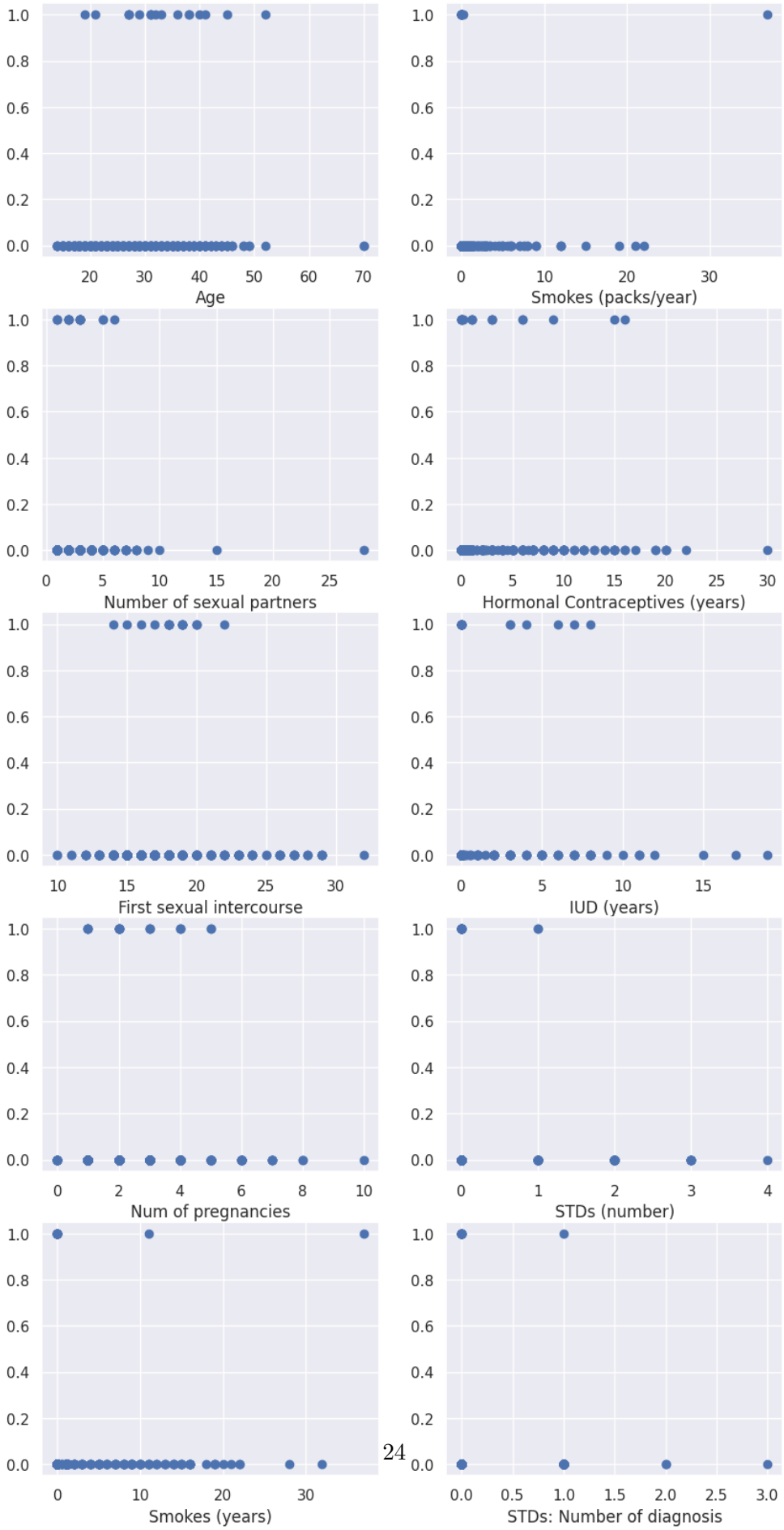
```
IUD                                  0.034055  0.087015  0.032660  0.046396
IUD (years)                         -0.006759  0.076929  0.009205  0.023429
STDs                                 0.058795  0.123132  0.055613  0.118630
STDs (number)                        0.075509  0.135198  0.063411  0.107235
STDs:condylomatosis                  0.060753  0.121458  0.067276  0.093542
STDs:vaginal condylomatosis         -0.014019 -0.020733 -0.015853 -0.017710
STDs:vulvo-perineal condylomatosis    0.062511  0.124405  0.069249  0.095947
STDs:syphilis                        0.011711  0.014714 -0.033911 -0.037884
STDs:pelvic inflammatory disease    -0.006997 -0.010348 -0.007913 -0.008839
STDs:genital herpes                 -0.006997 -0.010348 -0.007913  0.133093
STDs:molluscum contagiosum          -0.006997 -0.010348 -0.007913 -0.008839
STDs:HIV                             0.095108  0.132842  0.077943  0.131083
STDs:Hepatitis B                    -0.006997 -0.010348 -0.007913 -0.008839
STDs:HPV                            -0.009901 -0.014643 -0.011197 -0.012508
STDs: Number of diagnosis            0.079145  0.136406  0.056747  0.101398
Dx:Cancer                            0.136806  0.162374  0.115228  0.164877
Dx:CIN                              -0.019873  0.014644 -0.022473  0.126060
Dx:HPV                               0.136806  0.162374  0.115228  0.164877
Dx                                   0.077000  0.106942  0.093885  0.166946
Hinselmann                           1.000000  0.654458  0.199476  0.543296
Schiller                             0.654458  1.000000  0.357211  0.731103
Citology                             0.199476  0.357211  1.000000  0.317946
Biopsy                               0.543296  0.731103  0.317946  1.000000

[32 rows x 32 columns]
```

[384]:
```
x_corr = ccrf_df.corr().drop(num_cols, axis=0)
cat_corr = x_corr.drop(cat_cols, axis=1)

cat_corr

sns.set(rc = {'figure.figsize':(10, 15)})
sns.heatmap(cat_corr, annot=True, cmap='cividis',annot_kws={'size': 8},␣
 ↪square=True)
```

[384]:
```
<Axes: >
```


Smokes 0.039 0.23 -0.13 0.064 0.73 0.49 0.05 -0.047 0.1 0.092
Hormonal Contraceptives 0.12 0.019 0.027 0.17 0.023 0.014 0.4 -0.037 -0.038 -0.047
IUD 0.29 0.027 0.004 0.22 -0.00089 0.0023 0.16 0.74 0.06 0.035
STDs 0.037 0.052 -0.0032 0.064 0.1 0.029 0.0048 0.024 0.92 0.91
STDs:condylomatosis -0.0081 0.035 0.034 -0.03 0.05 0.0077 0.015 0.035 0.9 0.7
STDs:vaginal condylomatosis 0.012 -0.043 0.074 0.0015 0.12 0.042 -0.033 -0.0072 0.34 0.21
STDs:vulvo-perineal condylomatosis -0.0057 0.037 0.038 -0.029 0.052 0.0089 0.017 0.034 0.89 0.69
STDs:syphilis 0.023 0.027 -0.097 0.16 0.017 -0.0037 0.0024 -0.027 0.31 0.41
STDs:pelvic inflammatory disease 0.028 0.031 -2.9e-05 -0.056 -0.01 -0.0069 -0.012 -0.0083 0.054 0.1
STDs:genital herpes -0.029 -0.032 0.025 -0.031 -0.01 -0.0069 -0.017 -0.0083 0.054 0.1
STDs:molluscum contagiosum 0.0018 0.031 -0.012 0.045 -0.01 -0.0069 -0.019 -0.0083 0.054 0.1
STDs:HIV 0.0097 0.017 -0.0089 0.024 0.097 0.054 -0.036 0.018 0.39 0.55
STDs:Hepatitis B -0.029 -0.011 0.012 -0.031 0.11 0.1 -0.019 -0.0083 0.12 0.1
STDs:HPV 0.046 0.014 0.035 -0.026 0.055 -0.008 0.054 -0.012 0.077 0.066
Dx:Cancer 0.12 0.022 0.067 0.042 0.058 0.11 0.064 0.1 -0.012 -0.016
Dx:CIN 0.024 0.021 -0.018 0.02 -0.029 -0.02 0.01 0.01 -0.0057 0.012
Dx:HPV 0.11 0.027 0.044 0.054 0.061 0.11 0.084 0.037 -0.012 -0.016
Dx 0.077 0.026 0.047 0.031 -0.05 -0.034 0.011 0.1 -0.022 -0.00065
Hinselmann -0.017 -0.043 -0.017 0.037 0.028 0.019 0.057 -0.0068 0.076 0.079
Schiller 0.068 -0.011 -0.0018 0.066 0.046 0.013 0.085 0.077 0.14 0.14
Citology -0.025 0.024 -0.012 -0.022 -0.0026 0.0055 0.057 0.0092 0.063 0.057
Biopsy 0.042 -0.0028 0.0067 0.033 0.03 0.02 0.077 0.023 0.11 0.1
Age
Number of sexual partners
First sexual intercourse
Num of pregnancies
Smokes (years)
Smokes (packs/year)
Hormonal Contraceptives (years)
IUD (years)
STDs (number)
STDs: Number of diagnosis
0.8
0.6
0.4
0.2
0.0

[385]:
```
num1 = num_cols[:5]
num2 = num_cols[5:]

fig, ax = plt.subplots(nrows= 5, ncols=2, figsize=(10, 20))

ind = 0

for row in ax:
  row[0].scatter(ccrf_df[num1[ind]], ccrf_df['Dx:Cancer'])
  row[0].set_xlabel(num1[ind])
  row[1].scatter(ccrf_df[num2[ind]], ccrf_df['Dx:Cancer'])
  row[1].set_xlabel(num2[ind])
  ind += 1
```


Age
Smokes (packs/year)
Number of sexual partners
Hormonal Contraceptives (years)
First sexual intercourse
IUD (years)
Num of pregnancies
STDs (number)
Smokes (years)
STDs: Number of diagnosis