Day - 18 ____________ #100DaysOfML

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

df = pd.read_csv('Datasets/kag_risk_factors_cervical_cancer.csv')
pd.set_option('display.max_columns', 500)

df.replace('?',np.nan,inplace  = True)
df.shape[0]
```

```
858
```

```
# null percentage
for i in df.columns:
    print( 'Null values in ', i ,'are',df[i].isnull().sum())
```

```
Null values in  Age are 0
Null values in  Number of sexual partners are 26
Null values in  First sexual intercourse are 7
Null values in  Num of pregnancies are 56
Null values in  Smokes are 13
Null values in  Smokes (years) are 13
Null values in  Smokes (packs/year) are 13
Null values in  Hormonal Contraceptives are 108
Null values in  Hormonal Contraceptives (years) are 108
Null values in  IUD are 117
Null values in  IUD (years) are 117
Null values in  STDs are 105
Null values in  STDs (number) are 105
Null values in  STDs:condylomatosis are 105
Null values in  STDs:cervical condylomatosis are 105
Null values in  STDs:vaginal condylomatosis are 105
Null values in  STDs:vulvo-perineal condylomatosis are 105
Null values in  STDs:syphilis are 105
Null values in  STDs:pelvic inflammatory disease are 105
Null values in  STDs:genital herpes are 105
Null values in  STDs:molluscum contagiosum are 105
Null values in  STDs:AIDS are 105
Null values in  STDs:HIV are 105
Null values in  STDs:Hepatitis B are 105
Null values in  STDs:HPV are 105
Null values in  STDs: Number of diagnosis are 0
Null values in  STDs: Time since first diagnosis are 787
Null values in  STDs: Time since last diagnosis are 787
Null values in  Dx:Cancer are 0
Null values in  Dx:CIN are 0
Null values in  Dx:HPV are 0
Null values in  Dx are 0
Null values in  Hinselmann are 0
Null values in  Schiller are 0
Null values in  Citology are 0
Null values in  Biopsy are 0
```

```
for i in df.columns:
    print(  i ,'Null Percent = ',int(df[i].isnull().sum()/df.shape[0]
*100) )
```

```
Age Null Percent =  0
Number of sexual partners Null Percent =  3
First sexual intercourse Null Percent =  0
Num of pregnancies Null Percent =  6
Smokes Null Percent =  1
Smokes (years) Null Percent =  1
Smokes (packs/year) Null Percent =  1
Hormonal Contraceptives Null Percent =  12
Hormonal Contraceptives (years) Null Percent =  12
IUD Null Percent =  13
IUD (years) Null Percent =  13
STDs Null Percent =  12
STDs (number) Null Percent =  12
STDs:condylomatosis Null Percent =  12
STDs:cervical condylomatosis Null Percent =  12
STDs:vaginal condylomatosis Null Percent =  12
STDs:vulvo-perineal condylomatosis Null Percent =  12
STDs:syphilis Null Percent =  12
STDs:pelvic inflammatory disease Null Percent =  12
STDs:genital herpes Null Percent =  12
STDs:molluscum contagiosum Null Percent =  12
STDs:AIDS Null Percent =  12
STDs:HIV Null Percent =  12
STDs:Hepatitis B Null Percent =  12
STDs:HPV Null Percent =  12
STDs: Number of diagnosis Null Percent =  0
STDs: Time since first diagnosis Null Percent =  91
STDs: Time since last diagnosis Null Percent =  91
Dx:Cancer Null Percent =  0
Dx:CIN Null Percent =  0
Dx:HPV Null Percent =  0
Dx Null Percent =  0
Hinselmann Null Percent =  0
Schiller Null Percent =  0
Citology Null Percent =  0
Biopsy Null Percent =  0
```

```python
# columns to be dropped
# STDs: Time since first diagnosis Null Percent =  91
# STDs: Time since last diagnosis Null Percent =  91

df.drop(columns= ['STDs: Time since first diagnosis','STDs: Time since
last diagnosis'],inplace = True)

df.info()
```

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

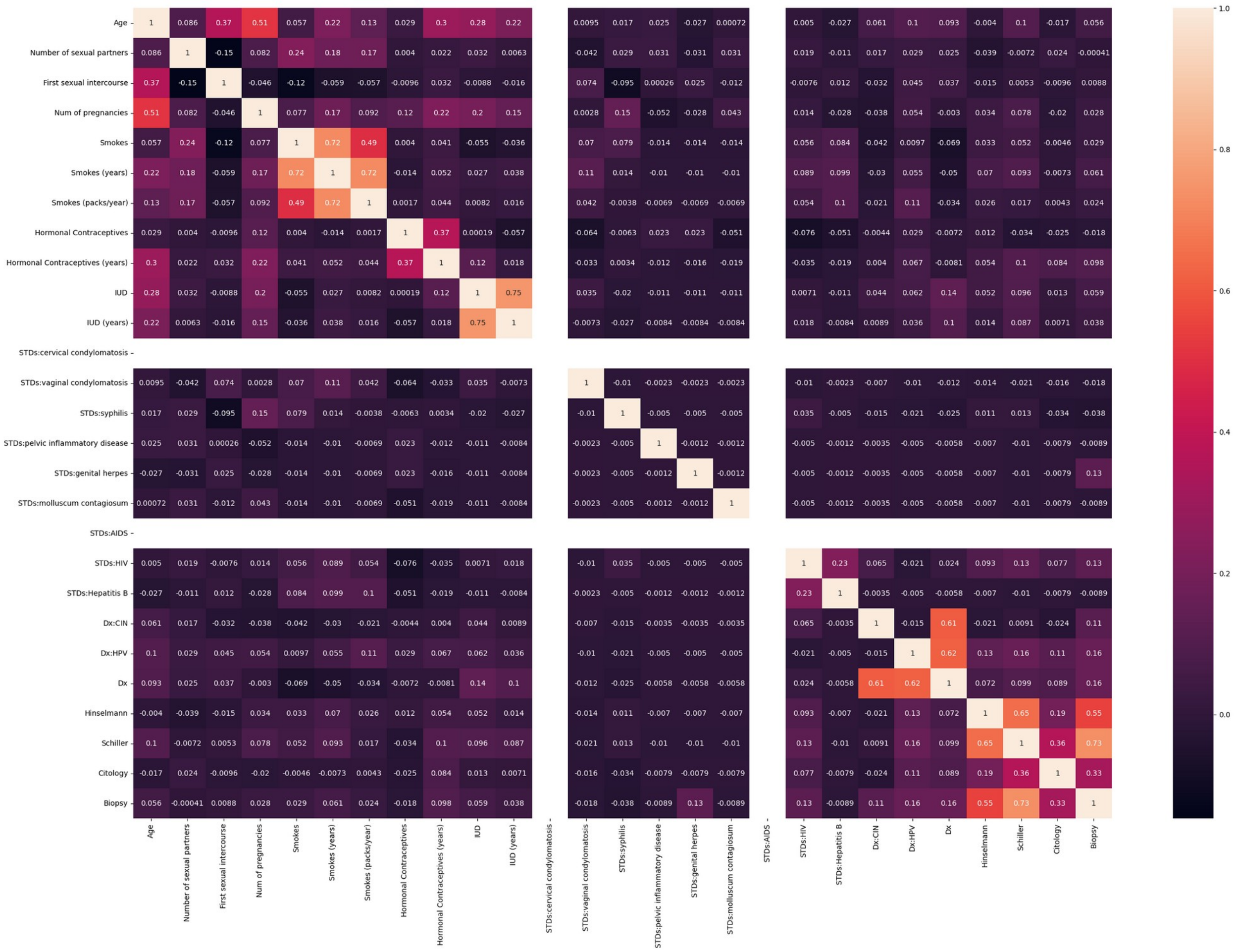
```
df['Age'].mean()
```

```
26.82051282051282
```

```
for i in df1.columns:
    df1[i] = df1[i].astype(float)
```

```
plt.figure(figsize=(30,20))
sns.heatmap(df1.corr(),annot= True)
```

```
<Axes: >
```



```
df1.drop(columns= ['STDs:condylomatosis','STDs:vulvo-perineal condylomatosis','STDs:condylomatosis'],inplace  = True)
```

```
df1.shape
```

```
(858, 27)
```

```
df1= df.copy()
```

```
X = df1.drop('Biopsy',axis =1)
y = df1['Biopsy']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.33, random_state=42)

from sklearn.metrics import accuracy_score

from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression(max_iter= 1000)

log_reg.fit(X_train_scaled,y_train)

log_reg_pred = log_reg.predict(X_test_scaled)
print(accuracy_score(y_test,log_reg_pred))
```

```
0.9577464788732394
```

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier()
clf.fit(X_train,y_train)
clf_pred = log_reg.predict(X_test)
print(accuracy_score(y_test,clf_pred))
```

```
0.9577464788732394
```

```
# Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train_scaled = sc.fit_transform(X_train)
X_test_scaled = sc.transform(X_test)

from sklearn.svm import SVC
svc = SVC()
svc.fit(X_train_scaled,y_train)
svc_pred = svc.predict(X_test_scaled)
print(accuracy_score(y_test,svc_pred))

final_pred = []
def voting(models, X_test):
    for i in X_test.index:
        ones=0
        zeroes = 0
        input1 = X_test[X_test.index == i].values
        for mod in models:
            out = mod.predict(input1)
            if out == 0:
                zeroes = zeroes+1
            else:
                ones = ones +1
        if(ones > zeroes):
            final_pred.append(1)
        else:
            final_pred.append(0)
    return final_pred
```

```
voting_out = voting([svc,log_reg,knn,naive],X_test)
```

```
len(voting_out)
```

```
284
```

```
X_train.shape,y_train.shape,X_test.shape,y_test.shape,df1.shape
```

```
((574, 26), (574,), (284, 26), (284,), (858, 27))
```

```
print(accuracy_score(y_test,voting_out))
```

```
0.9366197183098591
```

```
X_test[X_test.index == 713].values.shape
```

```
(1, 26)
```

```
X_test
```

```
      Age  Number of sexual partners  First sexual intercourse  \
713  16.0                        1.0                      16.0
604  23.0                        3.0                      17.0
120  33.0                        1.0                      16.0
208  27.0                        4.0                      16.0
380  18.0                        3.0                      15.0
..    ...                        ...                       ...
422  18.0                        2.0                      15.0
764  23.0                        1.0                      15.0
477  38.0                        2.0                      19.0
41   37.0                        2.0                      18.0
530  21.0                        4.0                      15.0

     Num of pregnancies  Smokes  Smokes (years)  Smokes
(packs/year)  \
713                 1.0     0.0             0.0                0.000

604                 2.0     0.0             0.0                0.000

120                 4.0     0.0             0.0                0.000

208                 1.0     0.0             0.0                0.000

380                 1.0     1.0             2.0                0.003
```

```
..                  ...     ...             ...                  ...

422                 2.0     1.0             0.5                0.050

764                 3.0     0.0             0.0                0.000

477                 2.0     0.0             0.0                0.000

41                  1.0     0.0             0.0                0.000

530                 1.0     0.0             0.0                0.000


     Hormonal Contraceptives  Hormonal Contraceptives (years)  IUD  \
713                      0.0                             0.00  0.0
604                      0.0                             0.00  0.0
120                      0.0                             0.00  0.0
208                      1.0                             0.67  0.0
380                      1.0                             0.58  0.0
..                       ...                              ...  ...
422                      1.0                             0.33  0.0
764                      1.0                             0.25  1.0
477                      0.0                             0.00  0.0
41                       0.0                             0.00  0.0
530                      0.0                             0.00  0.0

     IUD (years)  STDs:cervical condylomatosis  STDs:vaginal
condylomatosis  \
713          0.0                           0.0
0.0
604          0.0                           0.0
0.0
120          0.0                           0.0
0.0
208          0.0                           0.0
0.0
380          0.0                           0.0
0.0
..           ...                           ...
...
422          0.0                           0.0
0.0
764          7.0                           0.0
0.0
477          0.0                           0.0
0.0
41           0.0                           0.0
0.0
530          0.0                           0.0
0.0
```

```
     STDs:syphilis  STDs:pelvic inflammatory disease  STDs:genital
herpes  \
713            0.0                               0.0
0.0
604            0.0                               0.0
0.0
120            0.0                               0.0
0.0
208            0.0                               0.0
0.0
380            0.0                               0.0
0.0
..             ...                               ...
...
422            0.0                               0.0
0.0
764            0.0                               0.0
0.0
477            0.0                               0.0
0.0
41             0.0                               0.0
0.0
530            0.0                               0.0
0.0

     STDs:molluscum contagiosum  STDs:AIDS  STDs:HIV  STDs:Hepatitis B
\
713                         0.0        0.0       0.0               0.0

604                         0.0        0.0       0.0               0.0

120                         0.0        0.0       0.0               0.0

208                         0.0        0.0       0.0               0.0

380                         0.0        0.0       0.0               0.0

..                          ...        ...       ...               ...

422                         0.0        0.0       0.0               0.0

764                         0.0        0.0       0.0               0.0

477                         0.0        0.0       0.0               0.0

41                          0.0        0.0       1.0               0.0

530                         0.0        0.0       0.0               0.0
```

```
     Dx:CIN  Dx:HPV   Dx  Hinselmann  Schiller  Citology
713     0.0     0.0  0.0         0.0       0.0       0.0
604     0.0     0.0  0.0         0.0       0.0       0.0
120     0.0     0.0  0.0         0.0       0.0       0.0
208     0.0     0.0  0.0         0.0       0.0       0.0
380     0.0     0.0  0.0         0.0       0.0       0.0
..      ...     ...  ...         ...       ...       ...
422     0.0     0.0  0.0         1.0       0.0       0.0
764     0.0     0.0  0.0         0.0       0.0       0.0
477     0.0     0.0  0.0         0.0       0.0       0.0
41      1.0     0.0  1.0         0.0       1.0       0.0
530     0.0     0.0  0.0         0.0       1.0       1.0

[284 rows x 26 columns]
```

```
df
```

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

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

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

    STDs STDs (number) STDs:condylomatosis STDs:cervical
condylomatosis  \
0    0.0           0.0                 0.0
0.0
1    0.0           0.0                 0.0
0.0
2    0.0           0.0                 0.0
0.0
3    0.0           0.0                 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
0.0
856  0.0           0.0                 0.0
0.0
857  0.0           0.0                 0.0
0.0

    STDs:vaginal condylomatosis STDs:vulvo-perineal condylomatosis  \
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                                0.0
856                         0.0                                0.0
857                         0.0                                0.0

    STDs:syphilis STDs:pelvic inflammatory disease STDs:genital herpes
\
0             0.0                              0.0                 0.0

1             0.0                              0.0                 0.0

2             0.0                              0.0                 0.0

3             0.0                              0.0                 0.0

4             0.0                              0.0                 0.0

..            ...                              ...                 ...

853           0.0                              0.0                 0.0

854           0.0                              0.0                 0.0

855           0.0                              0.0                 0.0

856           0.0                              0.0                 0.0

857           0.0                              0.0                 0.0


    STDs:molluscum contagiosum STDs:AIDS STDs:HIV STDs:Hepatitis B
STDs:HPV  \
0                          0.0       0.0      0.0              0.0
0.0
1                          0.0       0.0      0.0              0.0
0.0
2                          0.0       0.0      0.0              0.0
0.0
3                          0.0       0.0      0.0              0.0
0.0
4                          0.0       0.0      0.0              0.0
0.0
..                         ...       ...      ...              ...
...
853                        0.0       0.0      0.0              0.0
0.0
854                        0.0       0.0      0.0              0.0
0.0
```

```
855                        0.0       0.0      0.0              0.0
0.0
856                        0.0       0.0      0.0              0.0
0.0
857                        0.0       0.0      0.0              0.0
0.0

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

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

[858 rows x 34 columns]
```

```
df[df.index == 9].values
```

```
array([[44, '3.0', '15.0', nan, '1.0', '1.266972909', '2.8', '0.0',
        '0.0', nan, nan, '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]], dtype=object)
```

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_train_scaled,y_train)
knn_pred = knn.predict(X_test_scaled)
print(accuracy_score(y_test,knn_pred))
# with scaled accurecy 0.9436619718309859
```

```
0.9401408450704225
```

```
from sklearn.naive_bayes import GaussianNB
naive = GaussianNB()
naive.fit(X_train,y_train)
naive_pred = naive.predict(X_test_scaled)
print(accuracy_score(y_test,naive_pred))
```

```
0.9366197183098591
```

```
C:\ProgramData\anaconda3\lib\site-packages\sklearn\base.py:420:
UserWarning: X does not have valid feature names, but GaussianNB was
fitted with feature names
  warnings.warn(
```

```
# voting_out = voting([svc,log_reg,knn,naive],X_test)
```

```
from sklearn.ensemble import RandomForestClassifier
rand_clf = RandomForestClassifier()
```

```
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
eclf1 = VotingClassifier(estimators=[('svc', svc), ('log_reg',
log_reg), ('knn', knn),('naive', naive),('random', rand_clf)],
voting='hard')
```

```
eclf1 = eclf1.fit(X_train_scaled, y_train)
```

```
from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression()
```

```
voting_pred = eclf1.predict(X_test_scaled)
```

```
print(accuracy_score(y_test,voting_pred))
```

```
0.9577464788732394
```