```
In [1]:
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sn
```

In [2]: path = 'https://raw.githubusercontent.com/ovibaridar/Data_sets/main/cancer%20pa

In [3]: data = pd.read_csv(path)

In [4]: data.head()

Out[4]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	
0	0	P1	33	1	2	4	5	4	3	2	
1	1	P10	17	1	3	1	5	3	4	2	
2	2	P100	35	1	4	5	6	5	5	4	
3	3	P1000	37	1	7	7	7	7	6	7	
4	4	P101	46	1	6	8	7	7	7	6	

5 rows × 26 columns

In [5]: data = data.drop(['index', 'Patient Id'], axis=1)

In [6]: data.head()

Out[6]:

	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity
0	33	1	2	4	5	4	3	2	2	4
1	17	1	3	1	5	3	4	2	2	2
2	35	1	4	5	6	5	5	4	6	7
3	37	1	7	7	7	7	6	7	7	7
4	46	1	6	8	7	7	7	6	7	7

5 rows × 24 columns

In [7]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 24 columns):

#	Column	Non-Null Count	Dtype
0	Age	1000 non-null	int64
1	Gender	1000 non-null	int64
2	Air Pollution	1000 non-null	int64
3	Alcohol use	1000 non-null	int64
4	Dust Allergy	1000 non-null	int64
5	OccuPational Hazards	1000 non-null	int64
6	Genetic Risk	1000 non-null	int64
7	chronic Lung Disease	1000 non-null	int64
8	Balanced Diet	1000 non-null	int64
9	Obesity	1000 non-null	int64
10	Smoking	1000 non-null	int64
11	Passive Smoker	1000 non-null	int64
12	Chest Pain	1000 non-null	int64
13	Coughing of Blood	1000 non-null	int64
14	Fatigue	1000 non-null	int64
15	Weight Loss	1000 non-null	int64
16	Shortness of Breath	1000 non-null	int64
17	Wheezing	1000 non-null	int64
18	Swallowing Difficulty	1000 non-null	int64
19	Clubbing of Finger Nails	1000 non-null	int64
20	Frequent Cold	1000 non-null	int64
21	Dry Cough	1000 non-null	int64
22	Snoring	1000 non-null	int64
23	Level	1000 non-null	object
d+\/n	oc. int(4/22) object(1)		-

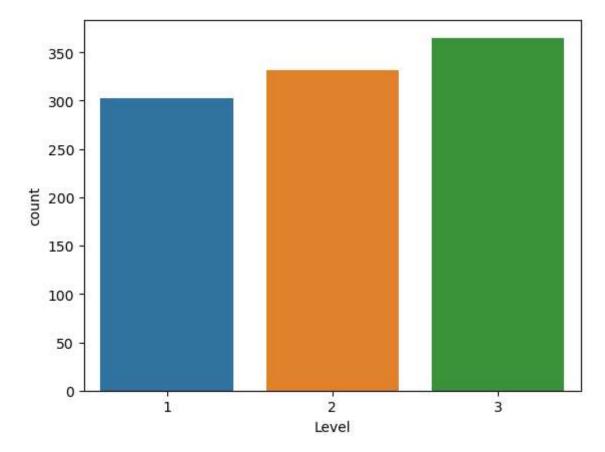
dtypes: int64(23), object(1)
memory usage: 187.6+ KB

```
In [8]: data.isnull().sum()
 Out[8]: Age
                                      0
                                       0
         Gender
         Air Pollution
                                       0
         Alcohol use
                                       0
         Dust Allergy
                                       0
         OccuPational Hazards
                                       0
                                       0
         Genetic Risk
         chronic Lung Disease
                                      0
         Balanced Diet
                                       0
         Obesity
                                       0
         Smoking
                                       0
                                       0
         Passive Smoker
         Chest Pain
                                       0
         Coughing of Blood
                                       0
                                       0
         Fatigue
         Weight Loss
                                       0
         Shortness of Breath
                                       0
         Wheezing
                                      0
         Swallowing Difficulty
                                       0
         Clubbing of Finger Nails
                                      0
         Frequent Cold
                                       0
                                       0
         Dry Cough
         Snoring
                                       0
         Level
                                       0
         dtype: int64
 In [9]: data.shape
 Out[9]: (1000, 24)
In [10]: data.columns
Out[10]: Index(['Age', 'Gender', 'Air Pollution', 'Alcohol use', 'Dust Allergy',
                 'OccuPational Hazards', 'Genetic Risk', 'chronic Lung Disease',
                 'Balanced Diet', 'Obesity', 'Smoking', 'Passive Smoker', 'Chest Pain',
                 'Coughing of Blood', 'Fatigue', 'Weight Loss', 'Shortness of Breath',
                 'Wheezing', 'Swallowing Difficulty', 'Clubbing of Finger Nails',
                 'Frequent Cold', 'Dry Cough', 'Snoring', 'Level'],
                dtype='object')
In [11]: | data['Level'] = data['Level'].replace({'Low': 1, 'Medium': 2, 'High': 3})
In [12]: data['Level'].value_counts()
Out[12]: Level
         3
               365
         2
               332
               303
         Name: count, dtype: int64
```

grafical view

```
In [13]: sn.countplot(data= data ,x='Level')
```

Out[13]: <Axes: xlabel='Level', ylabel='count'>

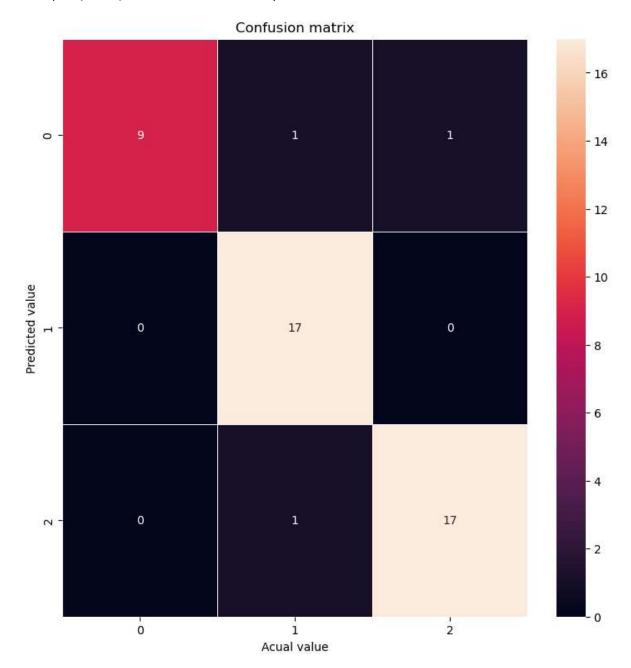


```
plt.figure(figsize=(10,10))
In [14]:
                 sn.heatmap(data.corr() ,annot = True)
Out[14]: <Axes: >
                                                                                                                                                            - 1.0
                                                  -0.20.099.15.03506207B.103004993407500499180532095.10.0335.095.10.0339.0130-02004706
                                                      0.250.230.20.190.230.210.10.130.250.180.230.150.142.0538043607360538403400653140.140.1
                                Air Pollution -.099.2: 1 0.750.640.610.710.630.520.60.480.610.590.610.210.260.200.055.080.240.170.26.020.6
                                Alcohol use -0.150.2-0.75 1 0.820.880.880.760.650.670.550.590.720.670.240.210.440.180.110.410.180.210.120.72
                                                                                                                                                            - 0.8
                                Dust Allergy -,0350.20.640.82 1 0.840.750.620.650.70.360.560.640.630.330.320.520.30.0310.350.220.30.056
                     OccuPational Hazards -,062.19.6 0.880.84 1 0.890.86 69.72 0.50.560.78 650.270.180.370.180.00 2030.070.16.02 6.0
                               Genetic Risk -.07-9.22.7 0.880.790.89 1 0.840.610.720.540.610.820.630.230.270.460.20.060.36.080.19.051
                      chronic Lung Disease -0.130.2 D.650.760.620.860.84 1 0.620.60.580.570.780.60.250.10.18.06.0070.0078.30.029.1 D.048.6
                                                                                                                                                            - 0.6
                             Balanced Diet -0049.10.520.650.650.690.680.62 1 0.710.650.730.80.750.40.006534.06040407040.260.330.150
                                              -.03<del>4</del>.120.60.670.70.720.730.60.71 1 0.490.680.670.810.550.310.410.094.130.150.290.20.030.83
                                               .07<del>5</del>0.2<mark>1</mark>0.480.550.360.50.540.580.650.49<mark>11</mark>0.760.650.560.2-0.20.02080407.240.040.040.010.190.5
                                    Smoking
                                              -0049.180.610.590.560.560.610.570.730.630.76<mark>1</mark>0.70.640.360.5080630.20.340.036.10.120.25
                            Passive Smoker
                                                                                                                                                            0.4
                                              -.01-30.220.590.720.6-40.780.830.780.80.670.650.7 1
                                                                                                    0.710.205.00101240.101.070208010419.140.140
                                              .05-10.150.610.670.630.650.630.60.750.810.560.640.71 1 0.480.110.340.08560805.0605.240.110.08 0.78
                         Coughing of Blood
                                     Fatigue -.090.120.210.240.330.270.230.250.40.550.20.380.250.48 1 0.470.40.170.16.0410.410.270.230
                                Weight Loss -0.1-0.05&260.210.320.180.270-0.0065310.20.0580010110.47 1 0.570.30.058.380.160.190.190.3
                                                                                                                                                            0.2
                       Shortness of Breath -.035.045.270.440.520.370.460.180.340.400.020068.240.320.40.57 1 0.21-0.20.470.350.490.160.5
                                  Wheezing -).09050706056.180.30.180.20.050706040904.0470.20.1-0.0805.170.330.21 1 0.390.390.0909054.120.2
                      Swallowing Difficulty -0.10.058080.10.0810009068003040.130.240.36.072086.16.0530.20.39 1 0.120.18.055.210.2
                                                                                                                                                            - 0.0
                   Clubbing of Finger Nails -.039.034.240.410.350.370.360.30.040.1-10.0411.036080.0660410.380.470.340.12 1 0.240.3-10.0168.24
                             Frequent Cold -).00100005.170.180.20.0707080029.260.290.040.10.048.240.410.160.35.099.130.24 1 0.520.340.4
                                 Dry Cough -.01-2.12.26.210.30.16.19.110.330.20.010.120.140.150.270.190.49.054.055.310.52 1 0.180.3
                                              -00477.168.020.112.0573.0737.0413.115.0319.190.250.119.0818.230.190.160.120.2-10.0162.340.118 1
                                                                                   Obesity
                                                                                                        Fatigue
                                               Age
                                                       Air Pollution
                                                           Alcohol use
                                                               Dust Allergy
                                                                   OccuPational Hazards
                                                                           chronic Lung Disease
                                                                               Balanced Diet
                                                                                        Smoking
                                                                                               Chest Pain
                                                                                                   Coughing of Blood
                                                                                                            Weight Loss
                                                                                                               Shortness of Breath
                                                                                                                        Swallowing Difficulty
                                                                                                                            Clubbing of Finger Nails
                                                                                                                                Frequent Cold
                                                                       Genetic Risk
                                                                                           Passive Smoker
                                                                                                                    Wheezing
                                                                                                                                    Dry Cough
                 data.duplicated().sum()
In [15]:
Out[15]: 848
In [16]:
                 data = data.drop_duplicates()
                 data.duplicated().sum()
Out[17]: 0
```

```
In [18]: data.shape
Out[18]: (152, 24)
In [19]: | x = data.drop('Level' , axis=1)
         y = data[['Level']]
In [20]: from sklearn.model_selection import train_test_split
In [21]: xtrain,xtest,ytrain,ytest = train_test_split(x,y ,random_state = 42 ,test_size
In [22]: xtrain.shape
Out[22]: (106, 23)
In [23]: xtest.shape
Out[23]: (46, 23)
In [24]: from sklearn.tree import DecisionTreeClassifier
In [25]: | dtc = DecisionTreeClassifier()
In [26]: dtc.fit(xtrain,ytrain)
Out[26]:
          ▼ DecisionTreeClassifier
          DecisionTreeClassifier()
In [27]: dtc.score(xtest,ytest)
Out[27]: 0.9347826086956522
In [28]: dtc.score(xtrain,ytrain)
Out[28]: 1.0
In [35]: from sklearn.metrics import accuracy score ,confusion matrix ,classification re
In [36]: | cm = confusion_matrix(ytest,dtc.predict(xtest))
         cm
Out[36]: array([[ 9, 1, 1],
                [ 0, 17, 0],
                [ 0, 1, 17]], dtype=int64)
```

```
In [37]: plt.figure(figsize=(9,9))
    sn.heatmap(cm,annot = True ,linewidth=.5)
    plt.xlabel('Acual value')
    plt.ylabel('Predicted value')
    plt.title('Confusion matrix')
```

Out[37]: Text(0.5, 1.0, 'Confusion matrix')



In [39]: print(classification_report(ytest,dtc.predict(xtest)))

support	f1-score	recall	precision	
11	0.90	0.82	1.00	1
17	0.94	1.00	0.89	2
18	0.94	0.94	0.94	3
46	0.93			accuracy
46	0.93	0.92	0.95	macro avg
46	0.93	0.93	0.94	weighted avg

In []: