lung-cancer-prediction-system

August 12, 2024

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objects as go
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
data = pd.read_csv("/content/survey lung cancer.csv")
data.head()
```

[2]:
```
  GENDER  AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
0      M   69        1               2        2              1
1      M   74        2               1        1              1
2      F   59        1               1        1              2
3      M   63        2               2        2              1
4      F   63        1               2        1              1

   CHRONIC DISEASE  FATIGUE   ALLERGY   WHEEZING  ALCOHOL CONSUMING  COUGHING  \
0                1         2         1         2                  2         2
1                2         2         2         1                  1         1
2                1         2         1         2                  1         2
3                1         1         1         1                  2         1
4                1         1         1         2                  1         2

   SHORTNESS OF BREATH  SWALLOWING DIFFICULTY  CHEST PAIN LUNG_CANCER
0                    2                      2           2         YES
1                    2                      2           2         YES
2                    2                      1           2          NO
3                    1                      2           2          NO
4                    2                      1           1          NO
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720ec8160>
```

```
from matplotlib import pyplot as plt
_df_0['AGE'].plot(kind='hist', bins=20, title='AGE')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_1['SMOKING'].plot(kind='hist', bins=20, title='SMOKING')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_2['YELLOW_FINGERS'].plot(kind='hist', bins=20, title='YELLOW_FINGERS')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_3['ANXIETY'].plot(kind='hist', bins=20, title='ANXIETY')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720ecb010>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
_df_4.groupby('GENDER').size().plot(kind='barh', color=sns.palettes.
 ↪mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
_df_5.groupby('LUNG_CANCER').size().plot(kind='barh', color=sns.palettes.
 ↪mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720ecfe50>
```

```
from matplotlib import pyplot as plt
_df_6.plot(kind='scatter', x='AGE', y='SMOKING', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_7.plot(kind='scatter', x='SMOKING', y='YELLOW_FINGERS', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_8.plot(kind='scatter', x='YELLOW_FINGERS', y='ANXIETY', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_9.plot(kind='scatter', x='ANXIETY', y='PEER_PRESSURE', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720ecacb0>
```

```
from matplotlib import pyplot as plt
_df_10['AGE'].plot(kind='line', figsize=(8, 4), title='AGE')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_11['SMOKING'].plot(kind='line', figsize=(8, 4), title='SMOKING')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_12['YELLOW_FINGERS'].plot(kind='line', figsize=(8, 4),␣
 ↪title='YELLOW_FINGERS')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_13['ANXIETY'].plot(kind='line', figsize=(8, 4), title='ANXIETY')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720eca6e0>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
plt.subplots(figsize=(8, 8))
df_2dhist = pd.DataFrame({
    x_label: grp['LUNG_CANCER'].value_counts()
    for x_label, grp in _df_14.groupby('GENDER')
})
sns.heatmap(df_2dhist, cmap='viridis')
plt.xlabel('GENDER')
_ = plt.ylabel('LUNG_CANCER')
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720ec9a50>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_15['GENDER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_15, x='AGE', y='GENDER', inner='stick', palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_16['LUNG_CANCER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_16, x='AGE', y='LUNG_CANCER', inner='stick', palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_17['GENDER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_17, x='SMOKING', y='GENDER', inner='stick', palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_18['LUNG_CANCER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_18, x='SMOKING', y='LUNG_CANCER', inner='stick',␣
 ↪palette='Dark2')
```

```
sns.despine(top=True, right=True, bottom=True, left=True)
```

[3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64
 2   SMOKING                309 non-null    int64
 3   YELLOW_FINGERS         309 non-null    int64
 4   ANXIETY                309 non-null    int64
 5   PEER_PRESSURE          309 non-null    int64
 6   CHRONIC DISEASE        309 non-null    int64
 7   FATIGUE                309 non-null    int64
 8   ALLERGY                309 non-null    int64
 9   WHEEZING               309 non-null    int64
 10  ALCOHOL CONSUMING      309 non-null    int64
 11  COUGHING               309 non-null    int64
 12  SHORTNESS OF BREATH    309 non-null    int64
 13  SWALLOWING DIFFICULTY  309 non-null    int64
 14  CHEST PAIN             309 non-null    int64
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

[4]:
```
data.head()
```

[4]:
```
  GENDER  AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
0      M   69        1               2        2              1
1      M   74        2               1        1              1
2      F   59        1               1        1              2
3      M   63        2               2        2              1
4      F   63        1               2        1              1

   CHRONIC DISEASE  FATIGUE   ALLERGY   WHEEZING  ALCOHOL CONSUMING  COUGHING  \
0                1        2         1         2                  2         2
1                2        2         2         1                  1         1
2                1        2         1         2                  1         2
3                1        1         1         1                  2         1
4                1        1         1         2                  1         2

   SHORTNESS OF BREATH  SWALLOWING DIFFICULTY  CHEST PAIN LUNG_CANCER
0                    2                      2           2         YES
1                    2                      2           2         YES
```

```
2                    2                      1           2          NO
3                    1                      2           2          NO
4                    2                      1           1          NO
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720b389a0>
```

```
from matplotlib import pyplot as plt
_df_35['AGE'].plot(kind='hist', bins=20, title='AGE')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_36['SMOKING'].plot(kind='hist', bins=20, title='SMOKING')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_37['YELLOW_FINGERS'].plot(kind='hist', bins=20, title='YELLOW_FINGERS')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_38['ANXIETY'].plot(kind='hist', bins=20, title='ANXIETY')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef72069f250>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
_df_39.groupby('GENDER').size().plot(kind='barh', color=sns.palettes.
 ↪mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
_df_40.groupby('LUNG_CANCER').size().plot(kind='barh', color=sns.palettes.
 ↪mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720c23820>
```

```
from matplotlib import pyplot as plt
_df_41.plot(kind='scatter', x='AGE', y='SMOKING', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_42.plot(kind='scatter', x='SMOKING', y='YELLOW_FINGERS', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_43.plot(kind='scatter', x='YELLOW_FINGERS', y='ANXIETY', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_44.plot(kind='scatter', x='ANXIETY', y='PEER_PRESSURE', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720aba470>
```

```
from matplotlib import pyplot as plt
_df_45['AGE'].plot(kind='line', figsize=(8, 4), title='AGE')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_46['SMOKING'].plot(kind='line', figsize=(8, 4), title='SMOKING')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_47['YELLOW_FINGERS'].plot(kind='line', figsize=(8, 4),␣
 ↪title='YELLOW_FINGERS')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_48['ANXIETY'].plot(kind='line', figsize=(8, 4), title='ANXIETY')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720b3b220>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
plt.subplots(figsize=(8, 8))
df_2dhist = pd.DataFrame({
    x_label: grp['LUNG_CANCER'].value_counts()
    for x_label, grp in _df_49.groupby('GENDER')
})
sns.heatmap(df_2dhist, cmap='viridis')
plt.xlabel('GENDER')
_ = plt.ylabel('LUNG_CANCER')
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720aba200>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_50['GENDER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_50, x='AGE', y='GENDER', inner='stick', palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_51['LUNG_CANCER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_51, x='AGE', y='LUNG_CANCER', inner='stick', palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_52['GENDER'].unique()))
```

```
plt.figure(figsize=figsize)
sns.violinplot(_df_52, x='SMOKING', y='GENDER', inner='stick', palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(_df_53['LUNG_CANCER'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(_df_53, x='SMOKING', y='LUNG_CANCER', inner='stick',␣
 ↪palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

[5]:
```
data.describe().T
```

[5]:
```
                       count       mean       std   min   25%   50%   75%  \
AGE                    309.0  62.673139  8.210301  21.0  57.0  62.0  69.0
SMOKING                309.0   1.563107  0.496806   1.0   1.0   2.0   2.0
YELLOW_FINGERS         309.0   1.569579  0.495938   1.0   1.0   2.0   2.0
ANXIETY                309.0   1.498382  0.500808   1.0   1.0   1.0   2.0
PEER_PRESSURE          309.0   1.501618  0.500808   1.0   1.0   2.0   2.0
CHRONIC DISEASE        309.0   1.504854  0.500787   1.0   1.0   2.0   2.0
FATIGUE                309.0   1.673139  0.469827   1.0   1.0   2.0   2.0
ALLERGY                309.0   1.556634  0.497588   1.0   1.0   2.0   2.0
WHEEZING               309.0   1.556634  0.497588   1.0   1.0   2.0   2.0
ALCOHOL CONSUMING      309.0   1.556634  0.497588   1.0   1.0   2.0   2.0
COUGHING               309.0   1.579288  0.494474   1.0   1.0   2.0   2.0
SHORTNESS OF BREATH    309.0   1.640777  0.480551   1.0   1.0   2.0   2.0
SWALLOWING DIFFICULTY  309.0   1.469256  0.499863   1.0   1.0   1.0   2.0
CHEST PAIN             309.0   1.556634  0.497588   1.0   1.0   2.0   2.0

                        max
AGE                    87.0
SMOKING                 2.0
YELLOW_FINGERS          2.0
ANXIETY                 2.0
PEER_PRESSURE           2.0
CHRONIC DISEASE         2.0
FATIGUE                 2.0
ALLERGY                 2.0
WHEEZING                2.0
ALCOHOL CONSUMING       2.0
COUGHING                2.0
SHORTNESS OF BREATH     2.0
SWALLOWING DIFFICULTY   2.0
CHEST PAIN              2.0
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef7207c5000>
```

```
from matplotlib import pyplot as plt
```

```
_df_19['mean'].plot(kind='hist', bins=20, title='mean')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_20['std'].plot(kind='hist', bins=20, title='std')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_21['min'].plot(kind='hist', bins=20, title='min')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_22['25%'].plot(kind='hist', bins=20, title='25%')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720a599f0>
```

```
from matplotlib import pyplot as plt
_df_23.plot(kind='scatter', x='mean', y='std', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_24.plot(kind='scatter', x='std', y='min', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_25.plot(kind='scatter', x='min', y='25%', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_26.plot(kind='scatter', x='25%', y='50%', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720ecb5e0>
```

```
from matplotlib import pyplot as plt
import seaborn as sns
def _plot_series(series, series_name, series_index=0):
  palette = list(sns.palettes.mpl_palette('Dark2'))
  xs = series['count']
  ys = series['mean']

  plt.plot(xs, ys, label=series_name, color=palette[series_index % len(palette)])

fig, ax = plt.subplots(figsize=(10, 5.2), layout='constrained')
df_sorted = _df_27.sort_values('count', ascending=True)
_plot_series(df_sorted, '')
sns.despine(fig=fig, ax=ax)
plt.xlabel('count')
_ = plt.ylabel('mean')
```

```
from matplotlib import pyplot as plt
import seaborn as sns
```

```
def _plot_series(series, series_name, series_index=0):
  palette = list(sns.palettes.mpl_palette('Dark2'))
  xs = series['count']
  ys = series['std']

  plt.plot(xs, ys, label=series_name, color=palette[series_index % len(palette)])

fig, ax = plt.subplots(figsize=(10, 5.2), layout='constrained')
df_sorted = _df_28.sort_values('count', ascending=True)
_plot_series(df_sorted, '')
sns.despine(fig=fig, ax=ax)
plt.xlabel('count')
_ = plt.ylabel('std')
```

```
from matplotlib import pyplot as plt
import seaborn as sns
def _plot_series(series, series_name, series_index=0):
  palette = list(sns.palettes.mpl_palette('Dark2'))
  xs = series['count']
  ys = series['min']

  plt.plot(xs, ys, label=series_name, color=palette[series_index % len(palette)])

fig, ax = plt.subplots(figsize=(10, 5.2), layout='constrained')
df_sorted = _df_29.sort_values('count', ascending=True)
_plot_series(df_sorted, '')
sns.despine(fig=fig, ax=ax)
plt.xlabel('count')
_ = plt.ylabel('min')
```

```
from matplotlib import pyplot as plt
import seaborn as sns
def _plot_series(series, series_name, series_index=0):
  palette = list(sns.palettes.mpl_palette('Dark2'))
  xs = series['count']
  ys = series['25%']

  plt.plot(xs, ys, label=series_name, color=palette[series_index % len(palette)])

fig, ax = plt.subplots(figsize=(10, 5.2), layout='constrained')
df_sorted = _df_30.sort_values('count', ascending=True)
_plot_series(df_sorted, '')
sns.despine(fig=fig, ax=ax)
plt.xlabel('count')
_ = plt.ylabel('25%')
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef72069e620>
```

```
from matplotlib import pyplot as plt
_df_31['mean'].plot(kind='line', figsize=(8, 4), title='mean')
```

```
plt.gca().spines[['top', 'right']].set_visible(False)

from matplotlib import pyplot as plt
_df_32['std'].plot(kind='line', figsize=(8, 4), title='std')
plt.gca().spines[['top', 'right']].set_visible(False)

from matplotlib import pyplot as plt
_df_33['min'].plot(kind='line', figsize=(8, 4), title='min')
plt.gca().spines[['top', 'right']].set_visible(False)

from matplotlib import pyplot as plt
_df_34['25%'].plot(kind='line', figsize=(8, 4), title='25%')
plt.gca().spines[['top', 'right']].set_visible(False)
```

[6]:

```
data.isna().sum()
```

[6]:

```
GENDER                   0
AGE                      0
SMOKING                  0
YELLOW_FINGERS           0
ANXIETY                  0
PEER_PRESSURE            0
CHRONIC DISEASE          0
FATIGUE                  0
ALLERGY                  0
WHEEZING                 0
ALCOHOL CONSUMING        0
COUGHING                 0
SHORTNESS OF BREATH      0
SWALLOWING DIFFICULTY    0
CHEST PAIN               0
LUNG_CANCER              0
dtype: int64
```

[7]:

```
data["LUNG_CANCER"].unique()
```

[7]:

```
array(['YES', 'NO'], dtype=object)
```

[8]:

```
data["GENDER"].unique()
```

[8]:

```
array(['M', 'F'], dtype=object)
```

[9]:

```
data["GENDER"] = data["GENDER"].map({'M': 2, 'F': 1 })
data['LUNG_CANCER'] = data['LUNG_CANCER'].map({'YES': 2, 'NO': 1 })
```

[10]:

```
data.dtypes
```

[10]:

```
GENDER                   int64
AGE                      int64
```

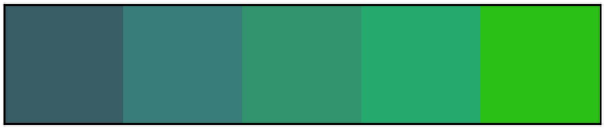
```
SMOKING                  int64
YELLOW_FINGERS           int64
ANXIETY                  int64
PEER_PRESSURE            int64
CHRONIC DISEASE          int64
FATIGUE                  int64
ALLERGY                  int64
WHEEZING                 int64
ALCOHOL CONSUMING        int64
COUGHING                 int64
SHORTNESS OF BREATH      int64
SWALLOWING DIFFICULTY    int64
CHEST PAIN               int64
LUNG_CANCER              int64
dtype: object
```

[11]:

```
def custom_palette(custom_colors):
    customPalette = sns.set_palette(sns.color_palette(custom_colors))
    sns.palplot(sns.color_palette(custom_colors),size=0.8)
    plt.tick_params(axis='both', labelsize=0,length = 0)
```

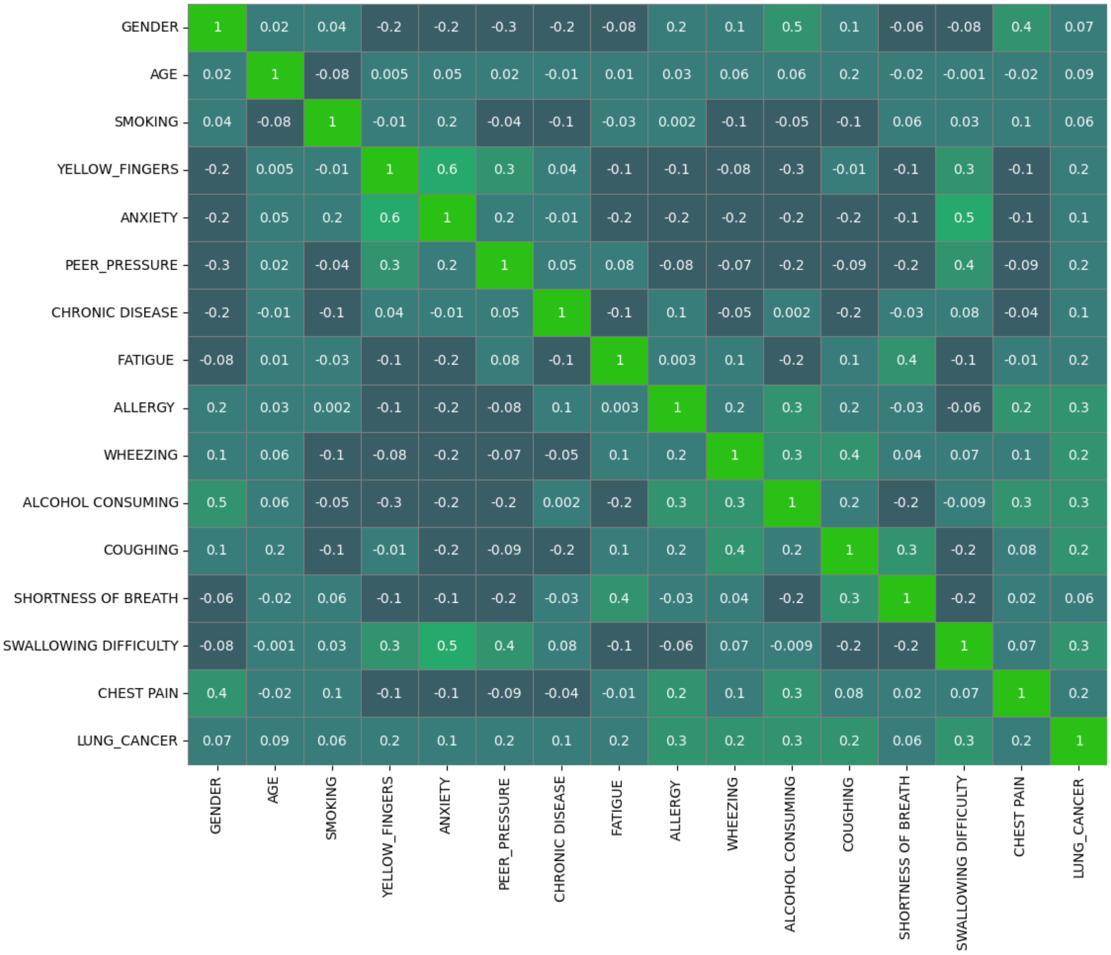
[12]:

```
pal = ["#395e66","#387d7a","#32936f","#26a96c","#2bc016"]
custom_palette(pal)
```



[15]:

```
fig, ax = plt.subplots(figsize=(12,10))
sns.heatmap(data.corr(),annot=True, fmt='.1g',cmap=pal, cbar=False,␣
 ↪linewidths=0.5, linecolor='grey');
```



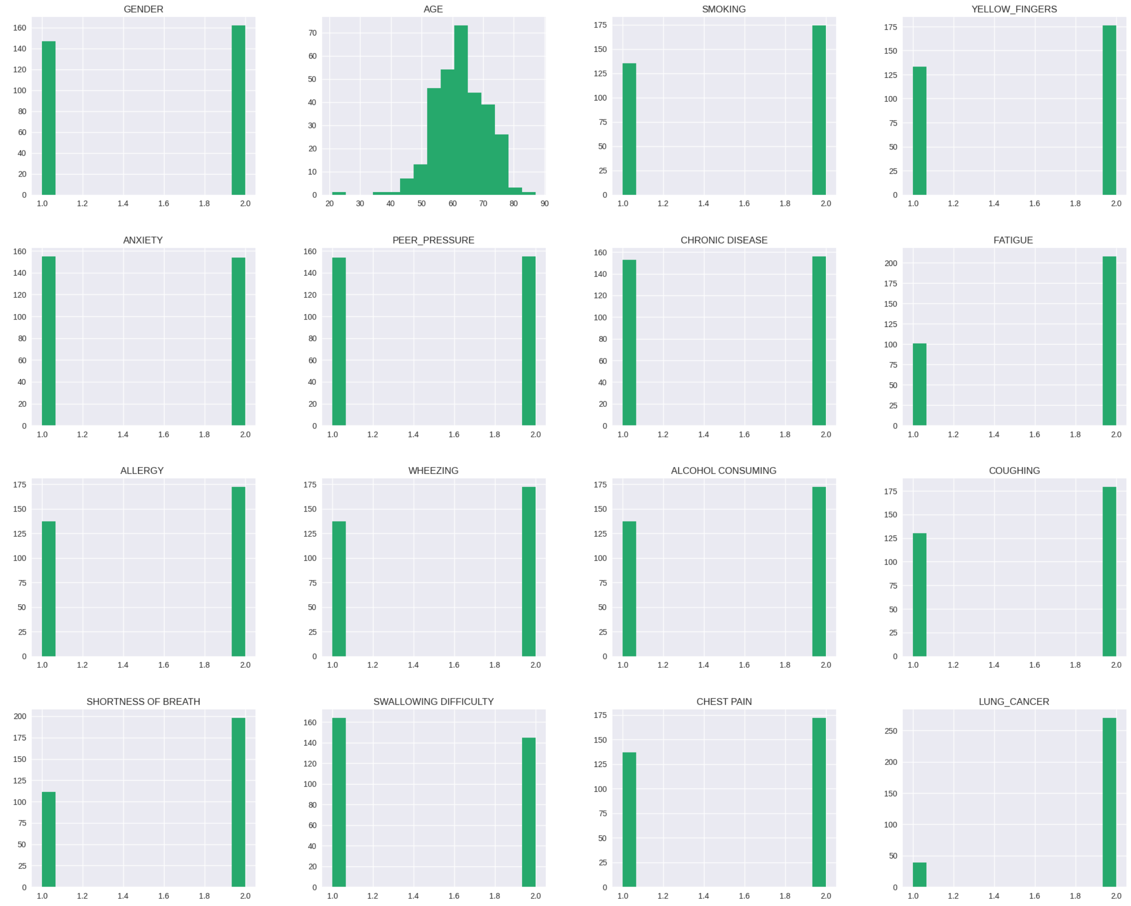
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
GENDER	1	0.02	0.04	-0.2	-0.2	-0.3	-0.2	-0.08	0.2	0.1	0.5	0.1	-0.06	-0.08	0.4	0.07
AGE	0.02	1	-0.08	0.005	0.05	0.02	-0.01	0.01	0.03	0.06	0.06	0.2	-0.02	-0.001	-0.02	0.09
SMOKING	0.04	-0.08	1	-0.01	0.2	-0.04	-0.1	-0.03	0.002	-0.1	-0.05	-0.1	0.06	0.03	0.1	0.06
YELLOW_FINGERS	-0.2	0.005	-0.01	1	0.6	0.3	0.04	-0.1	-0.1	-0.08	-0.3	-0.01	-0.1	0.3	-0.1	0.2
ANXIETY	-0.2	0.05	0.2	0.6	1	0.2	-0.01	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	0.5	-0.1	0.1
PEER_PRESSURE	-0.3	0.02	-0.04	0.3	0.2	1	0.05	0.08	-0.08	-0.07	-0.2	-0.09	-0.2	0.4	-0.09	0.2
CHRONIC DISEASE	-0.2	-0.01	-0.1	0.04	-0.01	0.05	1	-0.1	0.1	-0.05	0.002	-0.2	-0.03	0.08	-0.04	0.1
FATIGUE	-0.08	0.01	-0.03	-0.1	-0.2	0.08	-0.1	1	0.003	0.1	-0.2	0.1	0.4	-0.1	-0.01	0.2
ALLERGY	0.2	0.03	0.002	-0.1	-0.2	-0.08	0.1	0.003	1	0.2	0.3	0.2	-0.03	-0.06	0.2	0.3
WHEEZING	0.1	0.06	-0.1	-0.08	-0.2	-0.07	-0.05	0.1	0.2	1	0.3	0.4	0.04	0.07	0.1	0.2
ALCOHOL CONSUMING	0.5	0.06	-0.05	-0.3	-0.2	-0.2	0.002	-0.2	0.3	0.3	1	0.2	-0.2	-0.009	0.3	0.3
COUGHING	0.1	0.2	-0.1	-0.01	-0.2	-0.09	-0.2	0.1	0.2	0.4	0.2	1	0.3	-0.2	0.08	0.2
SHORTNESS OF BREATH	-0.06	-0.02	0.06	-0.1	-0.1	-0.2	-0.03	0.4	-0.03	0.04	-0.2	0.3	1	-0.2	0.02	0.06
SWALLOWING DIFFICULTY	-0.08	-0.001	0.03	0.3	0.5	0.4	0.08	-0.1	-0.06	0.07	-0.009	-0.2	-0.2	1	0.07	0.3
CHEST PAIN	0.4	-0.02	0.1	-0.1	-0.1	-0.09	-0.04	-0.01	0.2	0.1	0.3	0.08	0.02	0.07	1	0.2
LUNG_CANCER	0.07	0.09	0.06	0.2	0.1	0.2	0.1	0.2	0.3	0.2	0.3	0.2	0.06	0.3	0.2	1

[16]:
```
print ('Total Healthy Patients : {} '.format(data.LUNG_CANCER.
 ↪value_counts()[1]))
print ('Total Suspected Patients : {} '.format(data.LUNG_CANCER.
 ↪value_counts()[2]))
```

```
Total Healthy Patients : 39
Total Suspected Patients : 270
```

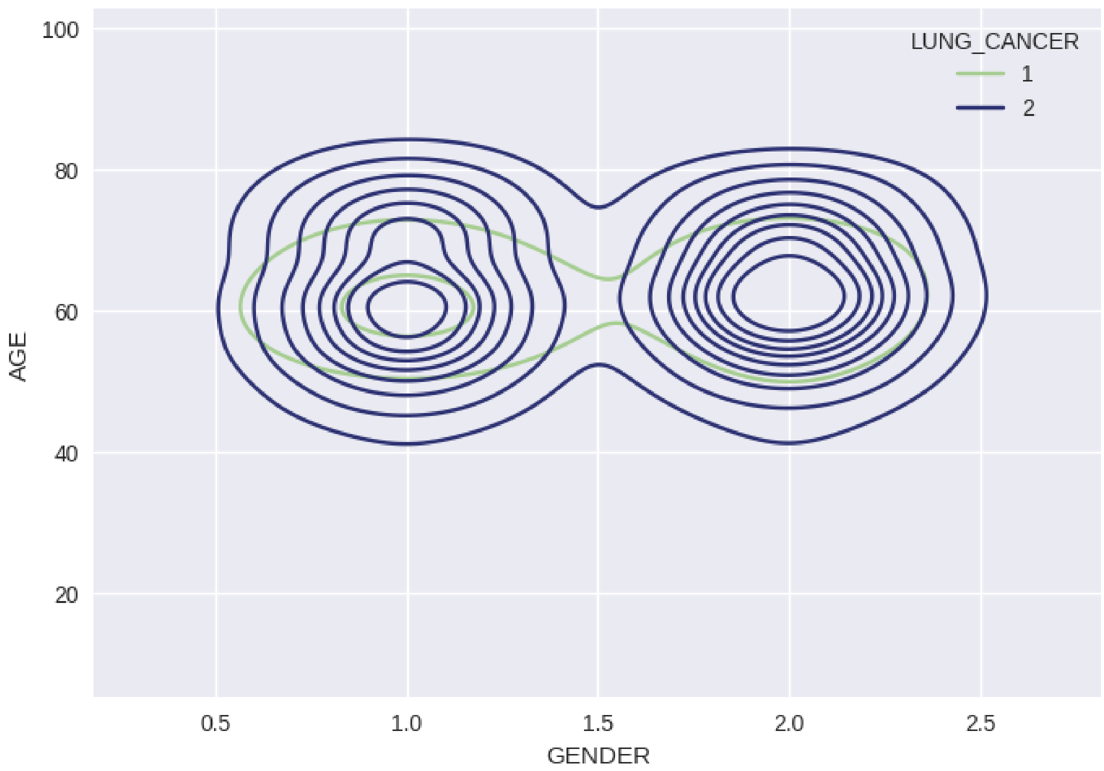
[18]:
```
values = data['LUNG_CANCER'].value_counts().tolist()
names = list(dict(data['LUNG_CANCER'].value_counts()).keys())

px.pie(data, values=values, names=names, hole = 0.5,
        color_discrete_sequence=["firebrick", "green"])
```

[19]:
```
plt.style.use("seaborn")
data.hist(figsize=(25,20), color=pal[3], bins=15);
```



[20]:

```
sns.kdeplot(x=data["GENDER"], y=data["AGE"], hue=data["LUNG_CANCER"], palette="crest");
plt.show()
```



Splitting and Training the data

```
[25]: X = data.drop(["LUNG_CANCER"], axis=1)
      X.head()
```

```
[25]:    GENDER  AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
      0       2   69        1               2        2              1
      1       2   74        2               1        1              1
      2       1   59        1               1        1              2
      3       2   63        2               2        2              1
      4       1   63        1               2        1              1

         CHRONIC DISEASE  FATIGUE   ALLERGY   WHEEZING  ALCOHOL CONSUMING  COUGHING  \
      0                1         2         1         2                  2         2
      1                2         2         2         1                  1         1
      2                1         2         1         2                  1         2
      3                1         1         1         1                  2         1
      4                1         1         1         2                  1         2

         SHORTNESS OF BREATH  SWALLOWING DIFFICULTY  CHEST PAIN
      0                    2                      2           2
      1                    2                      2           2
```

```
2                    2                      1           2
3                    1                      2           2
4                    2                      1           1
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef720b24bb0>
```

```
from matplotlib import pyplot as plt
_df_54['GENDER'].plot(kind='hist', bins=20, title='GENDER')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_55['AGE'].plot(kind='hist', bins=20, title='AGE')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_56['SMOKING'].plot(kind='hist', bins=20, title='SMOKING')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_57['YELLOW_FINGERS'].plot(kind='hist', bins=20, title='YELLOW_FINGERS')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef71feb1d80>
```

```
from matplotlib import pyplot as plt
_df_58.plot(kind='scatter', x='GENDER', y='AGE', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_59.plot(kind='scatter', x='AGE', y='SMOKING', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_60.plot(kind='scatter', x='SMOKING', y='YELLOW_FINGERS', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_61.plot(kind='scatter', x='YELLOW_FINGERS', y='ANXIETY', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7ef71ff02e90>
```

```
from matplotlib import pyplot as plt
_df_62['GENDER'].plot(kind='line', figsize=(8, 4), title='GENDER')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_63['AGE'].plot(kind='line', figsize=(8, 4), title='AGE')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_64['SMOKING'].plot(kind='line', figsize=(8, 4), title='SMOKING')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_65['YELLOW_FINGERS'].plot(kind='line', figsize=(8, 4),␣
 ↪title='YELLOW_FINGERS')
plt.gca().spines[['top', 'right']].set_visible(False)
```

[26]:

```
y = data["LUNG_CANCER"]
y.head()
```

[26]:

```
0    2
1    2
2    1
3    1
4    1
Name: LUNG_CANCER, dtype: int64
```

[31]:

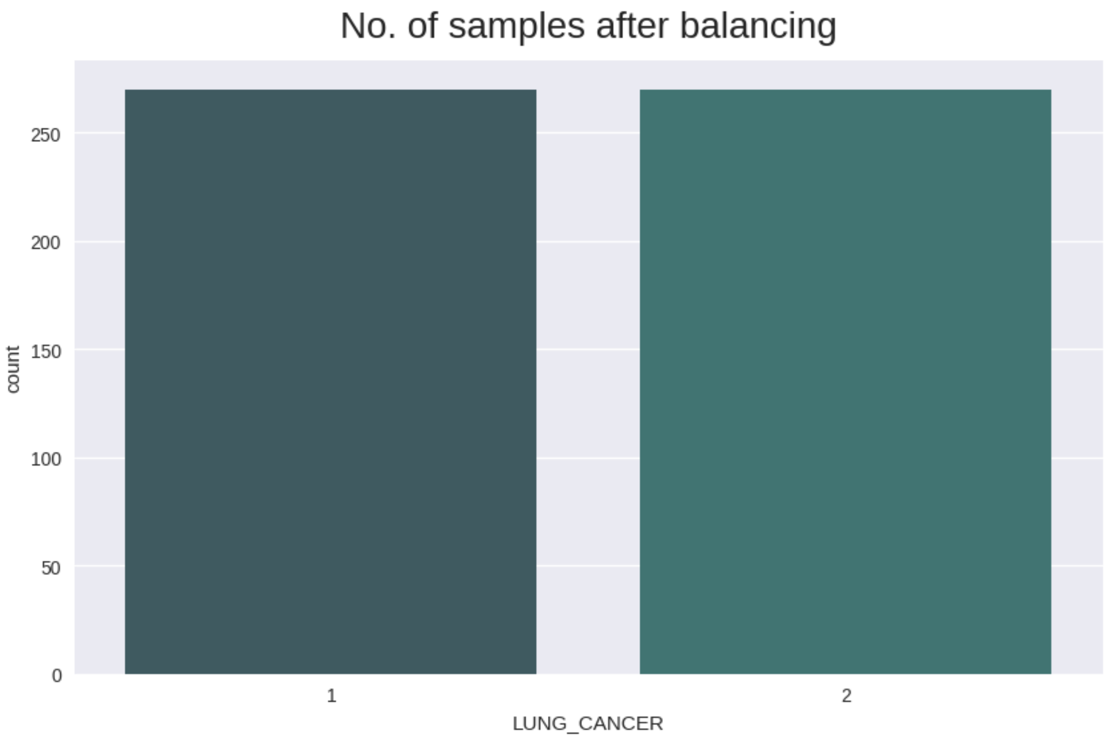
```
from imblearn.over_sampling import RandomOverSampler

over_samp = RandomOverSampler(random_state=0)
X_train_res, y_train_res = over_samp.fit_resample(X, y)
X_train_res.shape, y_train_res.shape
```

[31]:

```
((540, 15), (540,))
```

[32]:

```
plt.style.use("seaborn")
plt.figure(figsize=(10,6))
plt.title("No. of samples after balancing", fontsize=20, y=1.02)
sns.countplot(x = y_train_res, palette=pal)
plt.show()
```



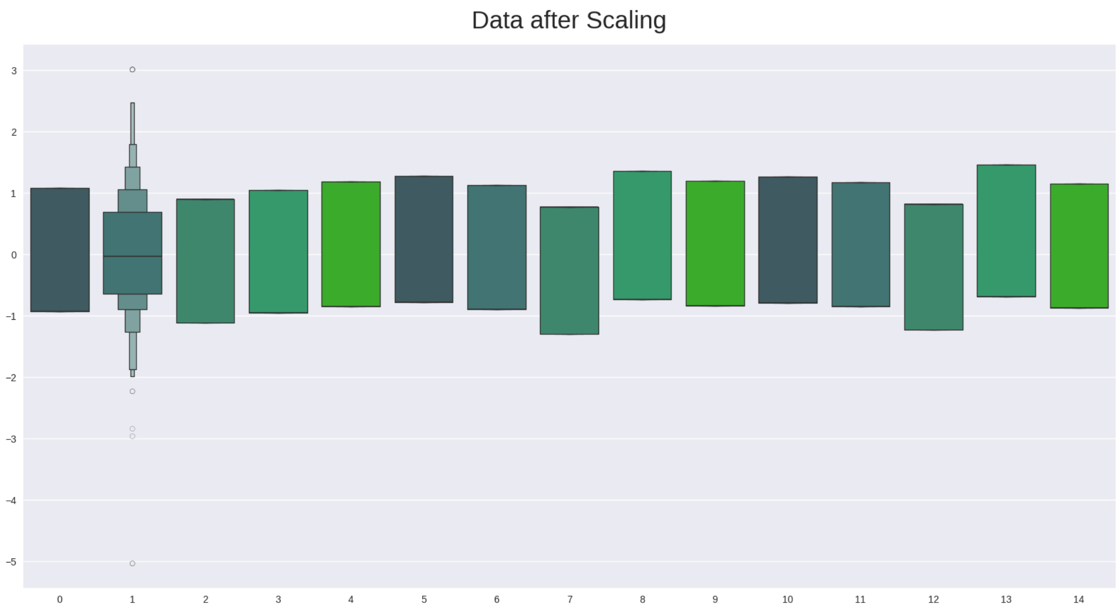
```
[33]: from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X_train_res, y_train_res,␣
       ↪test_size=0.2, random_state=42)
```

```
[34]: len(X_train), len(X_test)
```

```
[34]: (432, 108)
```

```
[35]: from sklearn.preprocessing import StandardScaler
      scaler = StandardScaler()
      X_train = scaler.fit_transform(X_train)
      X_test = scaler.transform(X_test)
```

```
[37]: plt.figure(figsize=(20,10))
      plt.title("Data after Scaling", fontsize=25, y=1.02)
      sns.boxenplot(data = X_train, palette=pal)
      plt.show()
```



Linear Regression

[38]:
```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(X_train, y_train)
```

[38]:
```
LinearRegression()
```

[40]:
```
LinearRegressionScore = lr.score(X_test, y_test)
print("Accuracy obtained by Linear Regression Score: ",␣
  ↪LinearRegressionScore*100)
```

```
Accuracy obtained by Linear Regression Score:  64.04214644616877
```

Decision Tree Classifier

[41]:
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
```

[41]:
```
DecisionTreeClassifier()
```

[42]:
```
DecisionTreeClassifierScore = dt.score(X_test, y_test)
print("Accuracy obtained by Decision Tree Classifier Score: ",␣
  ↪DecisionTreeClassifierScore*100)
```

```
Accuracy obtained by Decision Tree Classifier Score:  97.22222222222221
```