```
import numpy as np
import pandas as pd

data = pd.read_csv('datasets/lung cancer survey.csv')
data.head()
```

```
  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
```

```
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
```

```
data.isnull().sum()
```

```
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
```

```
data.shape
```

```
(309, 16)
```

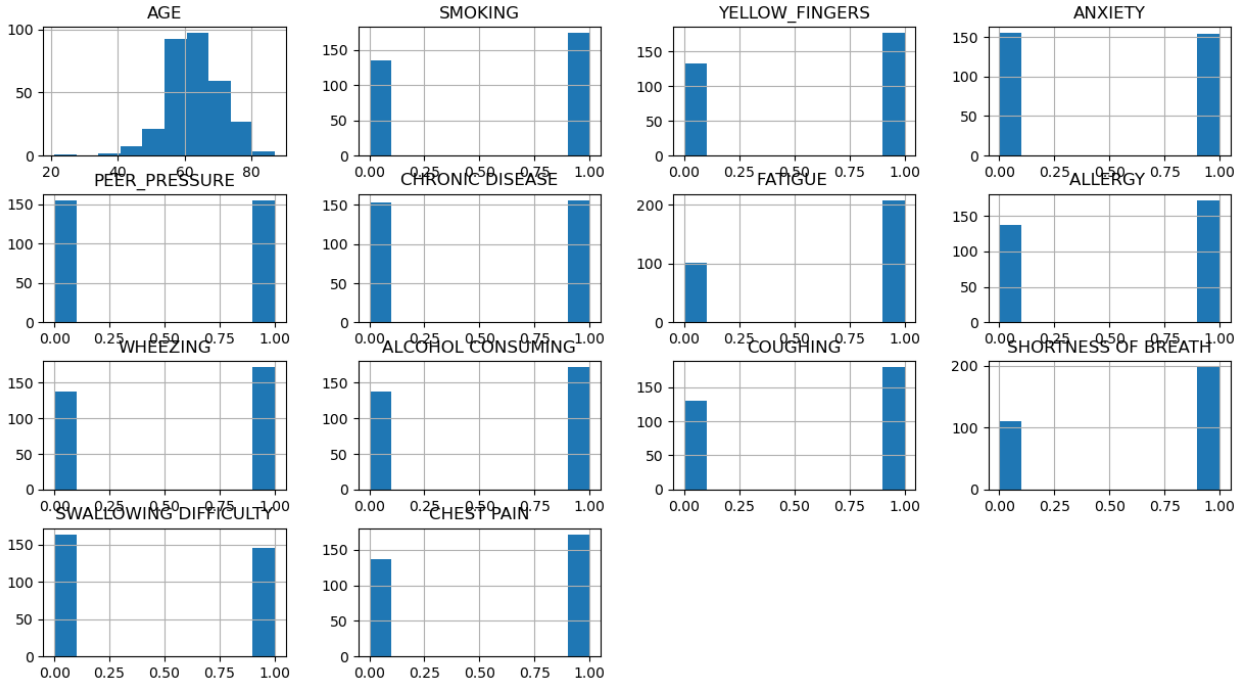
```
data = data.replace({1: 0, 2: 1})
```

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

```
data.hist(figsize=(15,8))
```

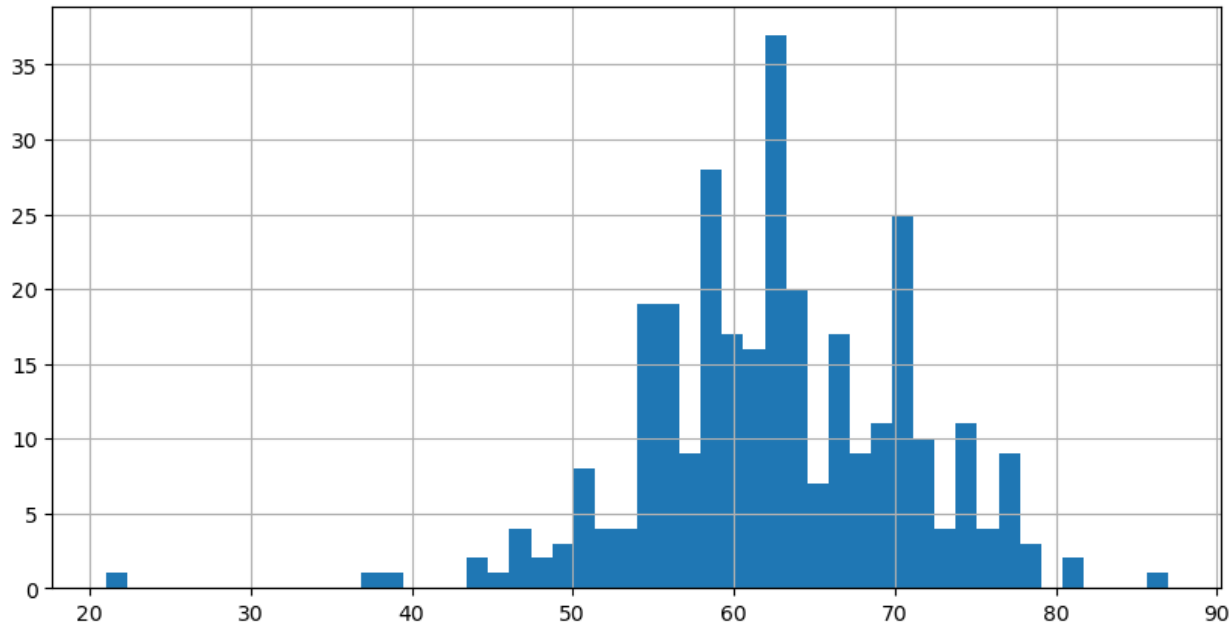
```
array([[<Axes: title={'center': 'AGE'}>,
        <Axes: title={'center': 'SMOKING'}>,
        <Axes: title={'center': 'YELLOW_FINGERS'}>,
        <Axes: title={'center': 'ANXIETY'}>],
       [<Axes: title={'center': 'PEER_PRESSURE'}>,
        <Axes: title={'center': 'CHRONIC DISEASE'}>,
        <Axes: title={'center': 'FATIGUE '}>,
```

```
       <Axes: title={'center': 'ALLERGY '}>],
      [<Axes: title={'center': 'WHEEZING'}>,
       <Axes: title={'center': 'ALCOHOL CONSUMING'}>,
       <Axes: title={'center': 'COUGHING'}>,
       <Axes: title={'center': 'SHORTNESS OF BREATH'}>],
      [<Axes: title={'center': 'SWALLOWING DIFFICULTY'}>,
       <Axes: title={'center': 'CHEST PAIN'}>, <Axes: >, <Axes: >]],
     dtype=object)
```

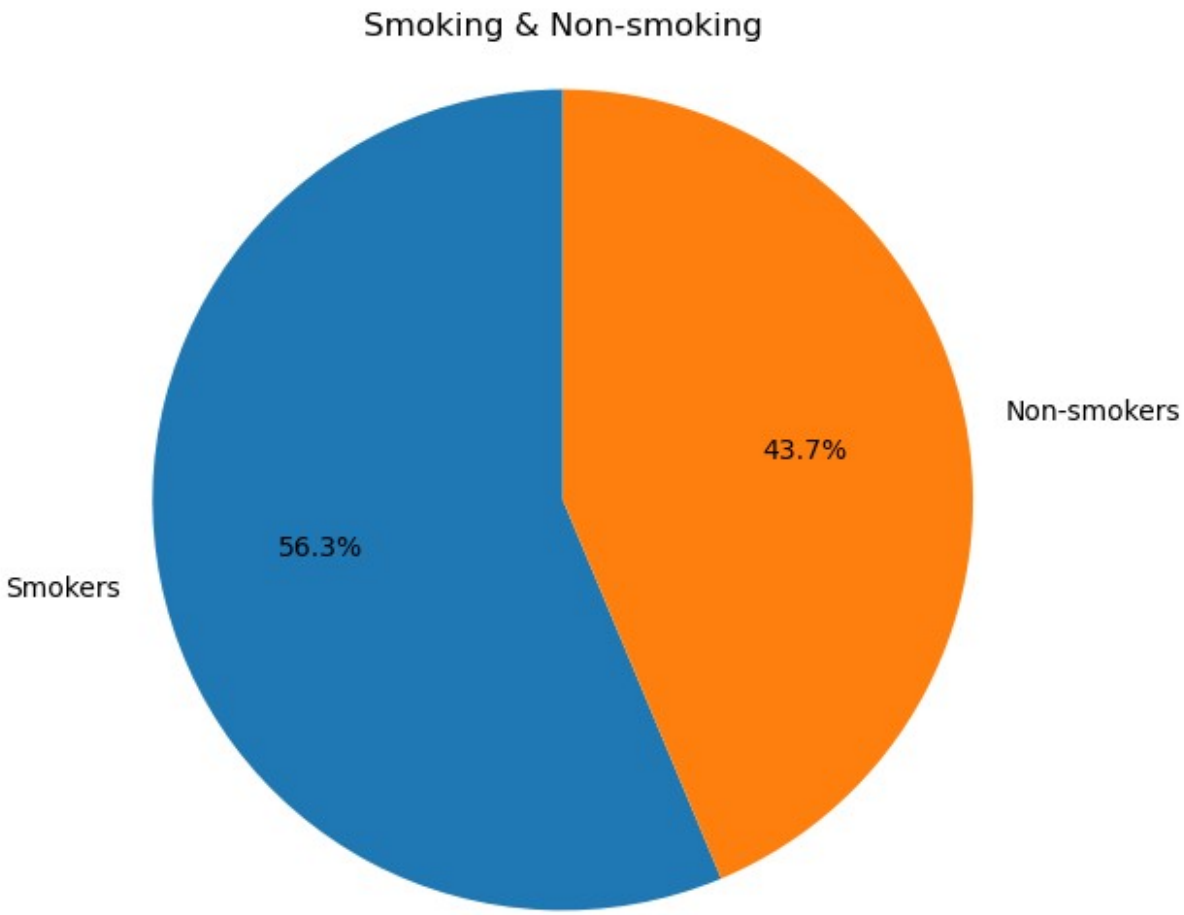



```
data['AGE'].hist(figsize=(10, 5), bins=50)
```

```
<Axes: >
```




```
smoking = data[data['SMOKING'] == 1].value_counts().sum()
non_smoking = data[data['SMOKING'] == 0].value_counts().sum()
labels = ['Smokers', 'Non-smokers']
sizes = [smoking, non_smoking]
plt.figure(figsize=(6, 6))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
plt.title('Smoking & Non-smoking')
plt.axis('equal')
plt.show()
```



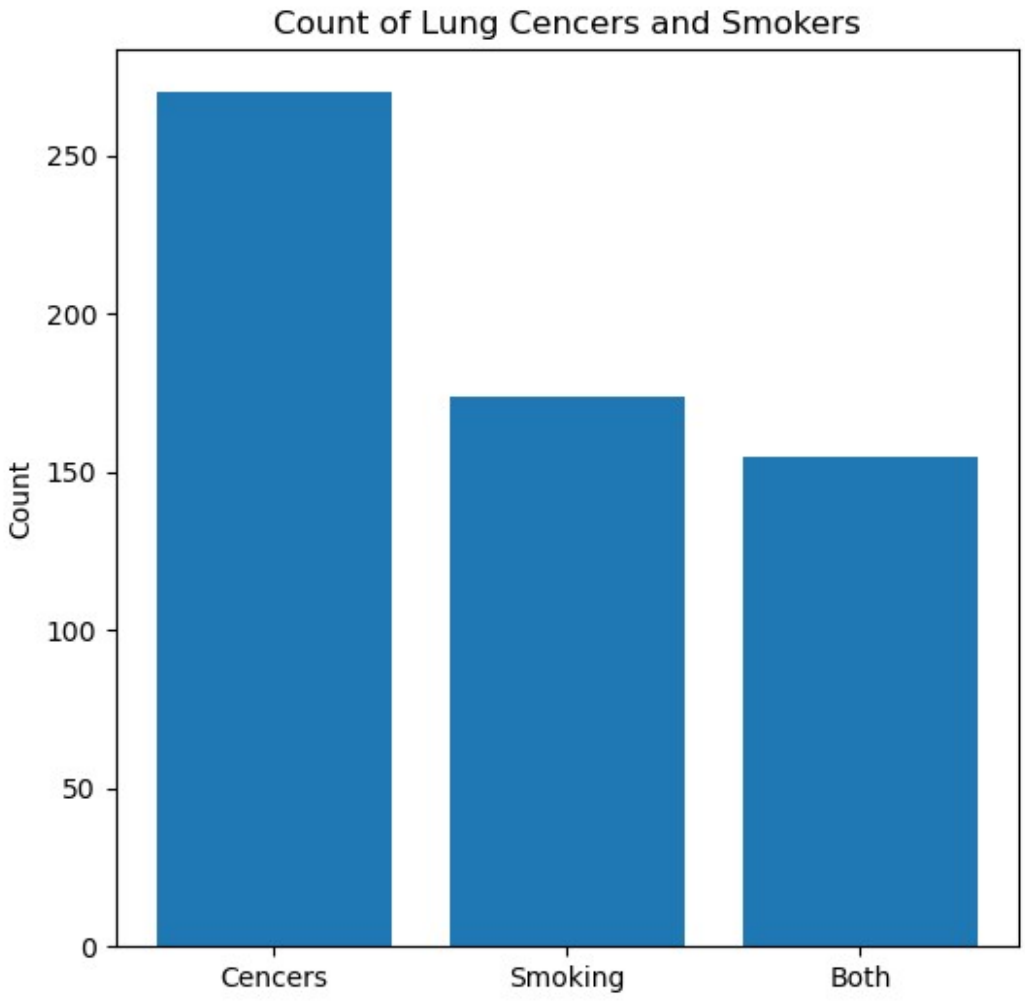

```
data['LUNG_CANCER'] = data['LUNG_CANCER'].replace({'YES':1, 'NO':0})
data.count()
```

```
/tmp/ipykernel_6615/962867787.py:1: FutureWarning: Downcasting
behavior in `replace` is deprecated and will be removed in a future
version. To retain the old behavior, explicitly call
`result.infer_objects(copy=False)`. To opt-in to the future behavior,
set `pd.set_option('future.no_silent_downcasting', True)`
  data['LUNG_CANCER'] = data['LUNG_CANCER'].replace({'YES':1, 'NO':0})
```

```
GENDER                  309
AGE                     309
SMOKING                 309
YELLOW_FINGERS          309
ANXIETY                 309
PEER_PRESSURE           309
CHRONIC DISEASE         309
FATIGUE                 309
ALLERGY                 309
WHEEZING                309
ALCOHOL CONSUMING       309
```

```
COUGHING                  309
SHORTNESS OF BREATH       309
SWALLOWING DIFFICULTY     309
CHEST PAIN                309
LUNG_CANCER               309
dtype: int64
```

```
cencers = data[(data['LUNG_CANCER'] == 1)]
smoking = data[(data['SMOKING'] == 1)]
both = data[(data['LUNG_CANCER'] == 1) & (data['SMOKING'] == 1)]
plt.figure(figsize=(6, 6))
plt.bar(['Cencers', 'Smoking', 'Both'], [cencers.shape[0],
smoking.shape[0], both.shape[0]])
plt.ylabel('Count')
plt.title('Count of Lung Cencers and Smokers')
plt.show()
```




```
data.corr(numeric_only=True)
```

```
                              AGE   SMOKING  YELLOW_FINGERS   ANXIETY  \
AGE                      1.000000 -0.084475        0.005205  0.053170
SMOKING                 -0.084475  1.000000       -0.014585  0.160267
YELLOW_FINGERS           0.005205 -0.014585        1.000000  0.565829
ANXIETY                  0.053170  0.160267        0.565829  1.000000
PEER_PRESSURE            0.018685 -0.042822        0.323083  0.216841
CHRONIC DISEASE         -0.012642 -0.141522        0.041122 -0.009678
FATIGUE                  0.012614 -0.029575       -0.118058 -0.188538
ALLERGY                  0.027990  0.001913       -0.144300 -0.165750
WHEEZING                 0.055011 -0.129426       -0.078515 -0.191807
ALCOHOL CONSUMING        0.058985 -0.050623       -0.289025 -0.165750
COUGHING                 0.169950 -0.129471       -0.012640 -0.225644
SHORTNESS OF BREATH     -0.017513  0.061264       -0.105944 -0.144077
SWALLOWING DIFFICULTY   -0.001270  0.030718        0.345904  0.489403
CHEST PAIN              -0.018104  0.120117       -0.104829 -0.113634
LUNG_CANCER              0.089465  0.058179        0.181339  0.144947

                         PEER_PRESSURE  CHRONIC DISEASE  FATIGUE
ALLERGY   \
AGE                           0.018685        -0.012642  0.012614
0.027990
SMOKING                      -0.042822        -0.141522 -0.029575
0.001913
YELLOW_FINGERS                0.323083         0.041122 -0.118058 -
0.144300
ANXIETY                       0.216841        -0.009678 -0.188538 -
0.165750
PEER_PRESSURE                 1.000000         0.048515  0.078148 -
0.081800
CHRONIC DISEASE               0.048515         1.000000 -0.110529
0.106386
FATIGUE                       0.078148        -0.110529  1.000000
0.003056
ALLERGY                      -0.081800         0.106386  0.003056
1.000000
WHEEZING                     -0.068771        -0.049967  0.141937
0.173867
ALCOHOL CONSUMING            -0.159973         0.002150 -0.191377
0.344339
COUGHING                     -0.089019        -0.175287  0.146856
0.189524
SHORTNESS OF BREATH          -0.220175        -0.026459  0.441745 -
0.030056
SWALLOWING DIFFICULTY         0.366590         0.075176 -0.132790 -
0.061508
CHEST PAIN                   -0.094828        -0.036938 -0.010832
0.239433
LUNG_CANCER                   0.186388         0.110891  0.150673
0.327766
```

```
                        WHEEZING  ALCOHOL CONSUMING  COUGHING  \
AGE                     0.055011           0.058985  0.169950
SMOKING                -0.129426          -0.050623 -0.129471
YELLOW_FINGERS         -0.078515          -0.289025 -0.012640
ANXIETY                -0.191807          -0.165750 -0.225644
PEER_PRESSURE          -0.068771          -0.159973 -0.089019
CHRONIC DISEASE        -0.049967           0.002150 -0.175287
FATIGUE                 0.141937          -0.191377  0.146856
ALLERGY                 0.173867           0.344339  0.189524
WHEEZING                1.000000           0.265659  0.374265
ALCOHOL CONSUMING       0.265659           1.000000  0.202720
COUGHING                0.374265           0.202720  1.000000
SHORTNESS OF BREATH     0.037834          -0.179416  0.277385
SWALLOWING DIFFICULTY   0.069027          -0.009294 -0.157586
CHEST PAIN              0.147640           0.331226  0.083958
LUNG_CANCER             0.249300           0.288533  0.248570

                        SHORTNESS OF BREATH  SWALLOWING DIFFICULTY
CHEST PAIN  \
AGE                           -0.017513              -0.001270   -
0.018104
SMOKING                        0.061264               0.030718
0.120117
YELLOW_FINGERS                -0.105944               0.345904   -
0.104829
ANXIETY                       -0.144077               0.489403   -
0.113634
PEER_PRESSURE                 -0.220175               0.366590   -
0.094828
CHRONIC DISEASE               -0.026459               0.075176   -
0.036938
FATIGUE                        0.441745              -0.132790   -
0.010832
ALLERGY                       -0.030056              -0.061508
0.239433
WHEEZING                       0.037834               0.069027
0.147640
ALCOHOL CONSUMING             -0.179416              -0.009294
0.331226
COUGHING                       0.277385              -0.157586
0.083958
SHORTNESS OF BREATH            1.000000              -0.161015
0.024256
SWALLOWING DIFFICULTY         -0.161015               1.000000
0.069027
CHEST PAIN                     0.024256               0.069027
1.000000
LUNG_CANCER                    0.060738               0.259730
```

```
0.190451
```

```
                        LUNG_CANCER
AGE                        0.089465
SMOKING                    0.058179
YELLOW_FINGERS             0.181339
ANXIETY                    0.144947
PEER_PRESSURE              0.186388
CHRONIC DISEASE            0.110891
FATIGUE                    0.150673
ALLERGY                    0.327766
WHEEZING                   0.249300
ALCOHOL CONSUMING          0.288533
COUGHING                   0.248570
SHORTNESS OF BREATH        0.060738
SWALLOWING DIFFICULTY      0.259730
CHEST PAIN                 0.190451
LUNG_CANCER                1.000000
```

```
X = data.iloc[:, 2:-1].values
y = data.iloc[:, -1].values
```

```
from sklearn.model_selection import train_test_split
```

```
from sklearn import metrics
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

```
import xgboost as xgb
```

```
xgb_clf = xgb.XGBClassifier()
xgb_clf.fit(X_train, y_train)
y_pred_xgb = xgb_clf.predict(X_test)
```

```
print(metrics.classification_report(y_test, y_pred_xgb))
f1_score_xgb = metrics.f1_score(y_test,y_pred_xgb)
```

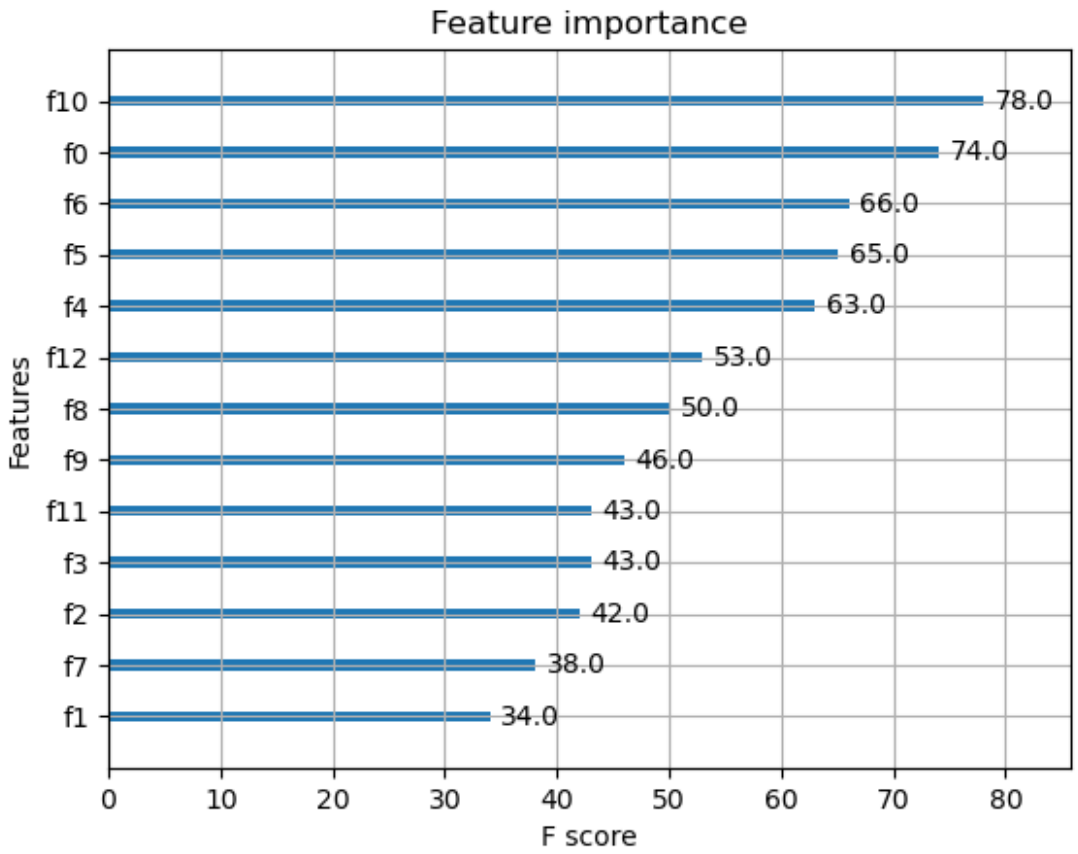
```
              precision    recall  f1-score   support

           0       0.80      0.50      0.62         8
           1       0.95      0.99      0.97        70

    accuracy                           0.94        78
   macro avg       0.87      0.74      0.79        78
weighted avg       0.93      0.94      0.93        78
```

```
xgb.plot_importance(xgb_clf)
```

```
<Axes: title={'center': 'Feature importance'}, xlabel='F score',
ylabel='Features'>
```




```
data.columns
```

```
Index(['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'],
      dtype='object')
```

```
from sklearn.neighbors import KNeighborsClassifier


knn_clf = KNeighborsClassifier(n_neighbors=10, leaf_size=60)
knn_clf.fit(X_train, y_train)
y_pred_knn = knn_clf.predict(X_test)
print(metrics.classification_report(y_test, y_pred_knn))
f1_score_knn = metrics.f1_score(y_test,y_pred_knn)
```

```
              precision    recall  f1-score   support

           0       0.75      0.38      0.50         8
           1       0.93      0.99      0.96        70

    accuracy                           0.92        78
```

```
   macro avg       0.84      0.68      0.73        78
weighted avg       0.91      0.92      0.91        78
```

```
from sklearn.svm import SVC

svc_clf = SVC(C=3, degree=5, kernel='linear')
svc_clf.fit(X_train, y_train)
y_pred_svc = svc_clf.predict(X_test)
print(metrics.classification_report(y_test, y_pred_svc))
f1_score_svc = metrics.f1_score(y_test,y_pred_svc)
```

```
              precision    recall  f1-score   support

           0       1.00      0.62      0.77         8
           1       0.96      1.00      0.98        70

    accuracy                           0.96        78
   macro avg       0.98      0.81      0.87        78
weighted avg       0.96      0.96      0.96        78
```

```
from sklearn.linear_model import LogisticRegression

log_reg = LogisticRegression(C=1.0, solver='lbfgs', max_iter=150)
log_reg.fit(X_train, y_train)
y_pred_lg = log_reg.predict(X_test)
print(metrics.classification_report(y_test, y_pred_lg))
f1_score_lg = metrics.f1_score(y_test,y_pred_lg)
```

```
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         8
           1       0.95      1.00      0.97        70

    accuracy                           0.95        78
   macro avg       0.97      0.75      0.82        78
weighted avg       0.95      0.95      0.94        78
```

```
data_1 = {
    'model' : ['LogisticRegression','KNN', 'SVC', 'XGBoost'],
    'f1_score': [f1_score_lg,f1_score_knn, f1_score_svc, f1_score_xgb]
}

F1 = pd.DataFrame(data_1)
F1.sort_values(by=['f1_score'],ascending=False)
```

```
                model  f1_score
2                 SVC  0.979021
0  LogisticRegression  0.972222
```

```
3              XGBoost  0.965035
1                  KNN  0.958333

# AmirHossein FeyzAbadi
```