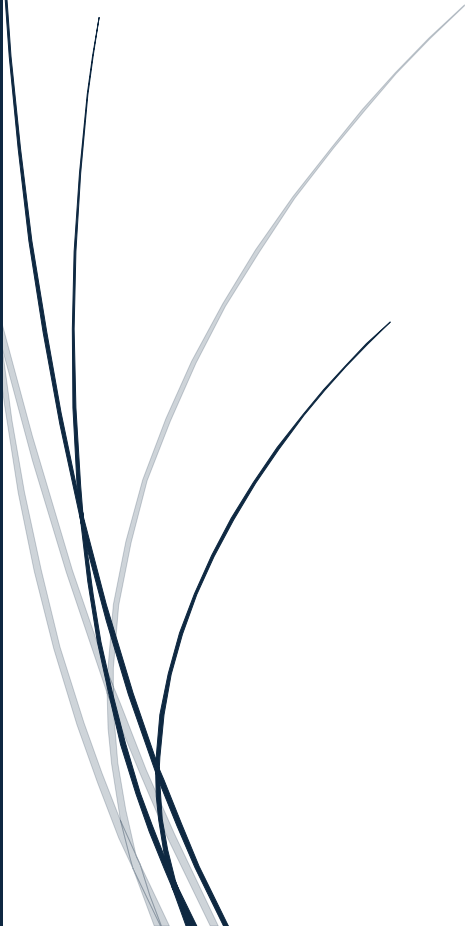


5/29/2025

PDAN8411

POE Part 2



Sven Kimi Masche
ST10030798

Contents

| | |
|--|----|
| Exploratory data analysis and classification model creation report:..... | 2 |
| Introduction: | 2 |
| Suitability of dataset: | 2 |
| Planning: | 2 |
| Exploratory data analysis results: | 4 |
| Data overview/cleaning: | 4 |
| Correlation analysis: | 6 |
| Univariate analysis: | 8 |
| Bivariate analysis: | 12 |
| Feature selection: | 21 |
| Model training: | 23 |
| References: | 29 |

Exploratory data analysis and classification model creation report:

Introduction:

This report covers the steps taken to explore and create a classification model based on a dataset featuring lung cancer status and its contributing factors. The goal is to explore the dataset and uncover any patterns, trends or information about the data that can help with the training of an accurate classification model that should be able to predict if a person is at risk of lung cancer.

Link to dataset:

<https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>

Suitability of dataset:

Before starting the process of exploratory data analysis and the creation of a classification model it's best to check if the data is suitable for classification algorithms. To conclude this, the following has been checked:

- **Balanced class distribution:** Classification algorithms work best when there is an equal balance of the predictive classes(encord, 2025). In this case the class "LUNG_CANCER" was imbalanced, with under 50 instances of "NO" and 240 instances of "YES". To combat this the dataset will go through SMOTE before being used in a classification model.
- **No multicollinearity.** This occurs when independent variables are too correlated with one another and can cause negative results in any classification model, such as overfitting (Bhandari, 2025). This has been tested for with a correlation matrix where none of the independent variables had a correlation coefficient over 0.8 with one another.

Planning:

Exploratory data analysis:

Data Overview/cleaning:

Before starting the full exploration analysis, we need to examine the data. By using panda's library, the plan is to load the csv file and check the various data types used in the data as well as to check for null values. For classification models and all forms of data analysis it's best to ensure all values are numerical, for instance where data is non-numerical, we can use label encoding or one-hot encoding to convert the data to numerical values.

Correlation analysis:

By using a heat map and pair-plots from the seaborn library, we can see how every feature in the data set interacts with one another. The pair plot will give a visual representation of how the features interact, and the heatmap will give us the numerical number representing the Pearson co-efficient. This tells us how much each pair of features affect one another. The purpose of this correlation analysis is to find the features that have the most impact on one another (Anon., 2025). For this case, we'll mostly be interested in how each feature interacts with the "LUNG_CANCER" feature, since that's what we are predicting.

Univariate analysis:

We will be looking at each individually to try and discern potential characteristics or patterns, such as outliers, skewedness and any other critical information that may help us in the bivariate analysis(Anon., 2024). The techniques to use may differ depending on the nature of the feature:

Categorical data and discrete numerical data:

For these types of data, we will have to make use of count plots to check the distribution of the data for each category.

Numerical data/discrete:

Here we can make use of boxplots and hist plots to try and uncover more about the data.

Bivariate analysis:

Comparisons between data features will start to be made to see how each feature impacts one another and perform a relationship analysis (Anon., 2024). The graphs that will primarily be used are histpots, box charts, and count plotsThe type of graph to be used as well as the combination of features to be checked against one another is dependent on the findings we gain as we perform the analysis.

Feature selection:

After exploratory analysis we will need to ensure only the most relevant features are used within the model. The form of feature selection to be used will be checking the p values of every feature and their effect on the feature that's desired to be predicted (in this case lung cancer). This will be done using sklearn's standard scaler import to create a stats model that can perform logistic regression using all features. The results are then displayed showing the p values of every independent variable. If the p-value is above 0.05 we don't use it (Banerjee Chandradip, 2023).

Model creation/evaluation:

With our selected features all that's left is the creation and evaluation of logistic regression models. For this report multiple logistic models are going to be created and have their outcomes checked against one another. It's planned to use a standard logistic regression model, KNN, decision tree, and random forest model.

Exploratory data analysis results:

Data overview/cleaning:

Using the panda's library, we can generate a brief overview of the data, as well as how many null values and the data types of each feature.

First five rows:

| | GENDER | AGE | SMOKING | YELLOW_FINGERS | ANXIETY | PEER_PRESSURE | CHRONIC DISEASE | FATIGUE | ALLERGY | WHEEZING | ALCOHOL CONSUMING | COUGHING | SHORTNESS OF BREATH | SWALLOWING DIFFICULTY |
|---|--------|-----|---------|----------------|---------|---------------|-----------------|---------|---------|----------|-------------------|----------|---------------------|-----------------------|
| 0 | M | 69 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 |
| 1 | M | 74 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 |
| 2 | F | 59 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 2 |
| 3 | M | 63 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| 4 | F | 63 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 |

| SWALLOWING DIFFICULTY | CHEST PAIN | LUNG_CANCER |
|-----------------------|------------|-------------|
| 2 | 2 | YES |
| 2 | 2 | YES |
| 1 | 2 | NO |
| 2 | 2 | NO |
| 1 | 1 | NO |

Null value count:

```
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 types:

```

GENDER          object
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     object
dtype: object

```

Here we can see that there are no null values, which is great. We can also see that all the features except for age are categorical. It also seems all categorical features have already been converted into numerical format except for LUNG_CANCER and GENDER. Those two features are to be label encoded as we need all values numerical.

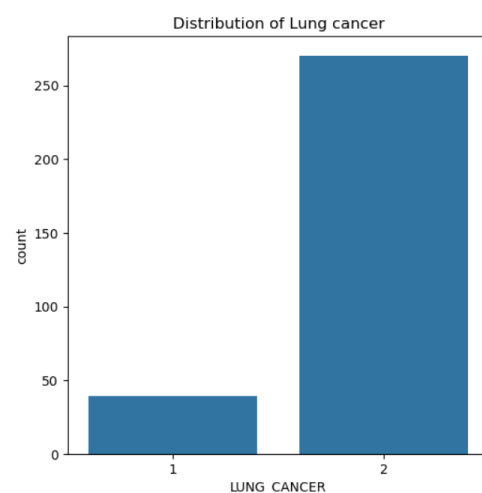
After label encoding our first few rows look like this:

| | GENDER | AGE | SMOKING | YELLOW_FINGERS | ANXIETY | PEER_PRESSURE | CHRONIC_DISEASE | FATIGUE | ALLERGY | WHEEZING | ALCOHOL_CONSUMING | COUGHING | SHORTNESS_OF_BREATH | SWALLOWING_DIFFICULTY |
|---|--------|-----|---------|----------------|---------|---------------|-----------------|---------|---------|----------|-------------------|----------|---------------------|-----------------------|
| 0 | 2 | 69 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 |
| 1 | 2 | 74 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 |
| 2 | 1 | 59 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 2 |
| 3 | 2 | 63 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| 4 | 1 | 63 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 |

| SWALLOWING_DIFFICULTY | CHEST_PAIN | LUNG_CANCER |
|-----------------------|------------|-------------|
| 2 | 2 | 2 |
| 2 | 2 | 2 |
| 1 | 2 | 1 |
| 2 | 2 | 1 |
| 1 | 1 | 1 |

It should also be noted that in lung_cancer, no and yes are 1 and 2 respectively, and in gender m and f are 1 and 2 respectively.

Unfortunately, we have the following issue with the data set:



This graph represents a huge class imbalance, meaning that almost all the data in the set is completely skewed. If we were to use this data to train a model, the model would most likely conclude that everything causes lung cancer.

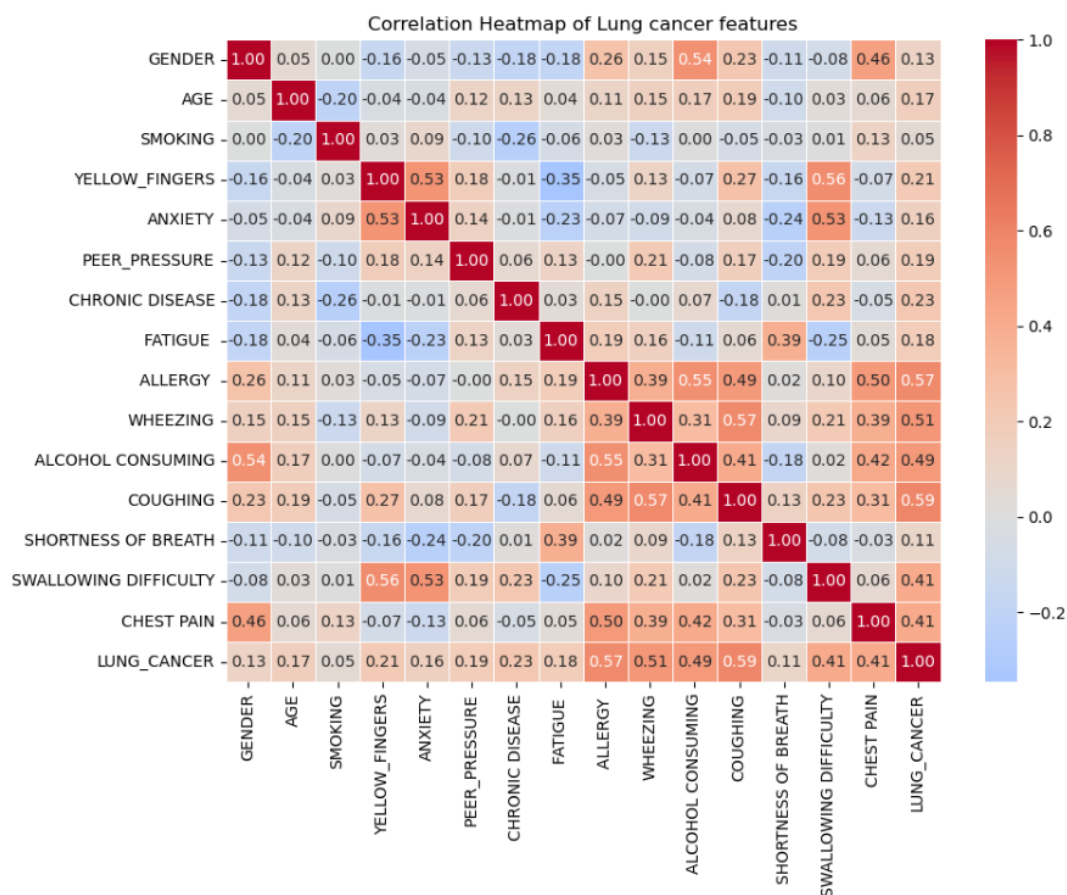
To combat this, two forms of preprocessing are done.

For the EDA we'd like to use real data, so the dataset has been downsized using panda's library to fix the class imbalance. This way we can rely on real data to find trends and patterns in the data. Unfortunately, it's downsized so far that it can't be used to accurately train a classification model. To combat this the data will be inflated using the SMOTE technique. The SMOTE version of the data frame will be used in model testing and training.

Correlation analysis:

Before we can individually analyze each of the features and their relationships with one another we should see how every feature is affected by one another. This will give us an idea of what patterns to look out for and where we should focus the most regarding the EDA. To check this a heatmap and a pair plot have been generated.

Heat map:



Each of the values within the heatmap represents the Pearson coefficient of the effect two features have on one another. Coefficient values between 0 and 1 usually show

significant affects, while values between -1 and 0 show non-significant effects (Anon., 2025). The purpose of this exercise is to try see what features have the most impact on the indication of lung cancer, so we can use these coefficients to help guide us to which features we should focus on the most.

The coefficients of each feature on lung cancer are:

Chest Pain: 0.41

Swallowing difficulty: 0.41

Shortness of breath: 0.11

Coughing: 0.59

Alcohol consuming: 0.49

Wheezing: 0.51

Allergy: 0.57

Fatigue: 0.18

Chronic disease: 0.23

Peer pressure: 0.19

Anxiety: 0.16

Yellow fingers: 0.21

Smoking: 0.05

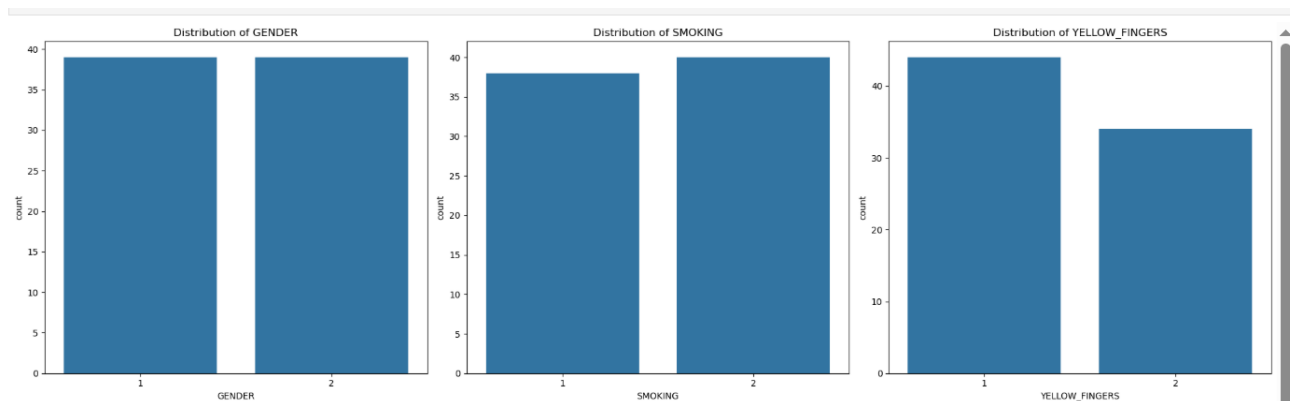
Age: 0.17

Gender: 0.13

Looking at these coefficients, unsurprisingly typical lung cancer symptoms such as coughing, wheezing, chest pain and swallowing difficulty are strongly associated with lung cancer, however surprisingly smoking, which is a proven carcinogen has the lowest effect on lung cancer, which doesn't make sense. This needs to be thoroughly investigated.

Univariate analysis:

This is part of the EDA in which we look at each individual feature. Most of the data is discrete being 1's and 2's ("NO" and "YES") so we'll heavily rely on count plots.



Gender:

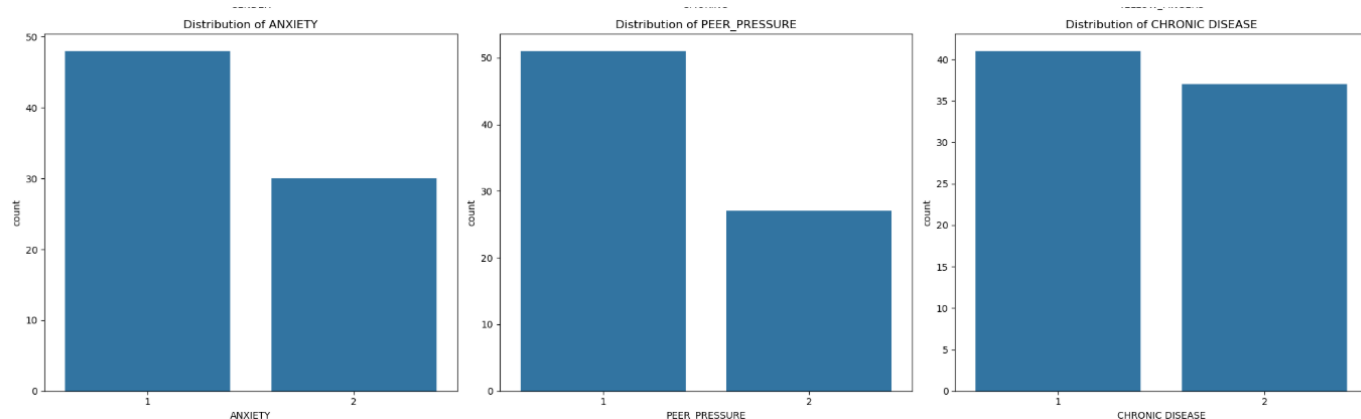
Gender is equally balanced, possibly an effect of downsizing the dataset.

Smoking:

The number of Smokers is slightly more than non-smokers. Considering smoking is a carcinogen and plays a large role in the cause of lung cancer we should keep a close eye on this feature(Walser et al., 2008).

Yellow fingers:

A substantially higher number of individuals without yellow fingers. Yellow fingers are indicative of heavy smokers, as it represents a higher frequency in smoking, to the point where it stains one's fingers(Northrup et al., 2022). This in turn should give it quite good predictive value, assuming that smoking is a carcinogen, more smoking should have an even higher chance of lung cancer.



Anxiety:

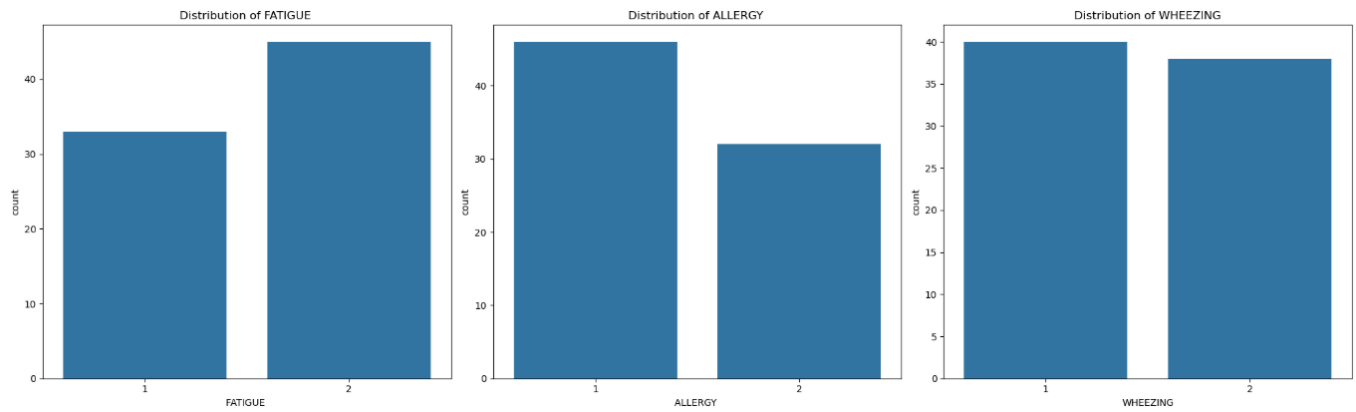
Much less people with anxiety compared to those without. Smoking and anxiety are medically linked to one another as it seems individuals who smoke have a higher chance of having anxiety (NHS, 2025). There may be some predictive potential.

Peer pressure:

Significantly less individuals experience peer pressure, which is also another factor that may influence an individual's smoking status as peer pressure is usually a cause of smoking or alcohol use (Leshargie et al., 2019).

Chronic diseases:

Slightly less people with chronic diseases. Chronic diseases share the same symptoms of lung cancer as well as can also be caused by smoking (Lee, Walser and Dubinett, 2009). We can see where its predictive power comes from, as individuals with chronic diseases in theory are likely to also have lung cancer.



Fatigue:

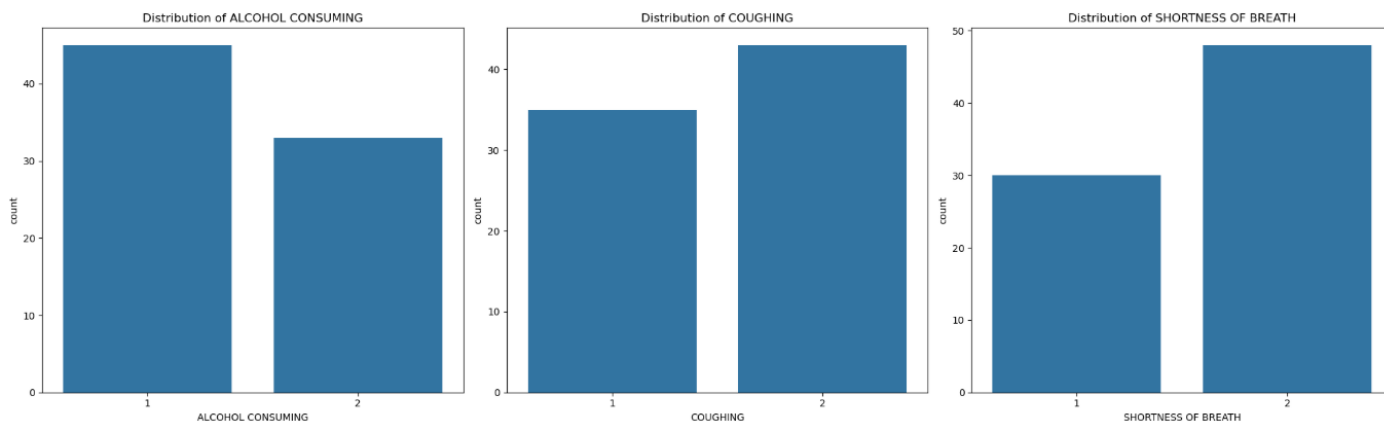
A substantial number of individuals experience fatigue. Fatigue is medically considered to be linked to lung cancer which proves its predictive power (Cleveland Clinic, 2025).

Allergy:

Substantially less people with allergies. There is no definitive link between lung cancer and allergies; in fact some studies indicate that allergies have a negative effect on the development of lung cancer (Khan et al., 2024).

Wheezing:

Slight decrease in the number of people who experience wheezing. Wheezing is a common symptom of lung cancer (healthline, 2025).



Alcohol consuming:

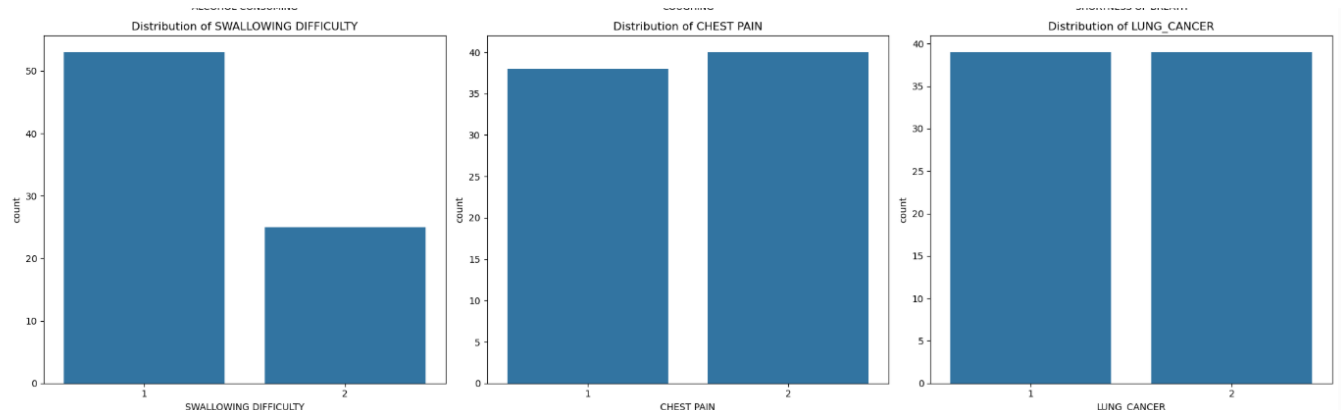
Substantially smaller number of individuals who consume alcohol. Alcohol is linked to lung cancer as a carcinogen. Alcohol destroys cells which increases the likelihood of developing any cancer, as well as individuals who consume alcohol are also likely to smoke (National Cancer Institute, 2025).

Coughing:

Slightly higher amount of people who cough. Coughing is yet another symptom of lung cancer and can prove to be a predictive value(healthline, 2025).

Shortness of breath:

Significant number of people have shortness of breath. Yet another lung cancer symptom with high predictive potential (healthline, 2025).



Swallowing difficulty:

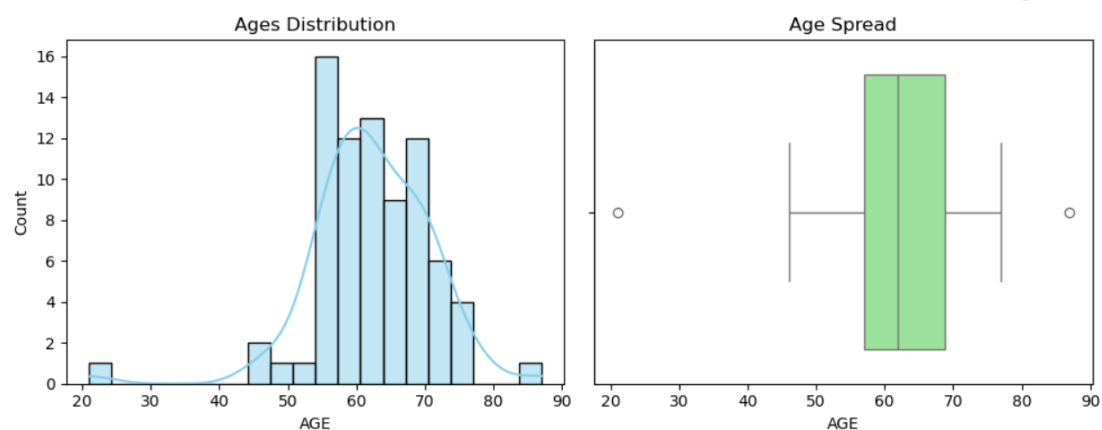
Significantly less people with swallowing difficulty. Difficulty swallowing is yet another lung cancer symptom, making it imperative to see its effect (healthline, 2025).

Chest pain:

slight increase in individuals with chest pain. Another lung cancer symptom and strong predictor(healthline, 2025).

Lung cancer:

Even distribution thanks to the down sampling of the dataset.



Age:

Most of the ages in the dataset are between 57 and 69 with few yet present outliers. Lung cancer is linked to higher ages so we can expect age to have some predictive potential.

Key take aways:

From the univariate analysis we can create a few different assortments of the features:

Smoking and alcohol consumption are the main carcinogens directly linked to lung cancer.

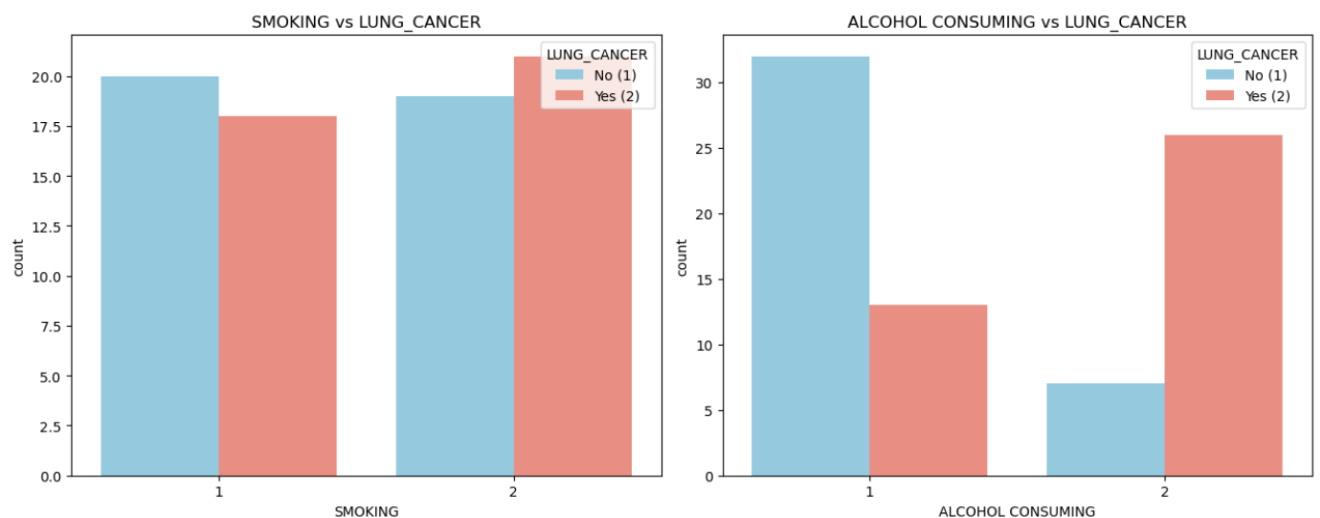
Anxiety, yellow fingertips, and peer pressure are contributing factors of carcinogens. Chest pain, swallowing difficulty, coughing, wheezing, shortness of breath, and fatigue are all symptoms of lung cancer.

Chronic illness and allergies are other conditions that may be indicative of lung cancer. Age and gender are demographic considerations which need to be considered.

Bivariate analysis:

We can now look at the deeper relationships between the different features.

Main carcinogens:



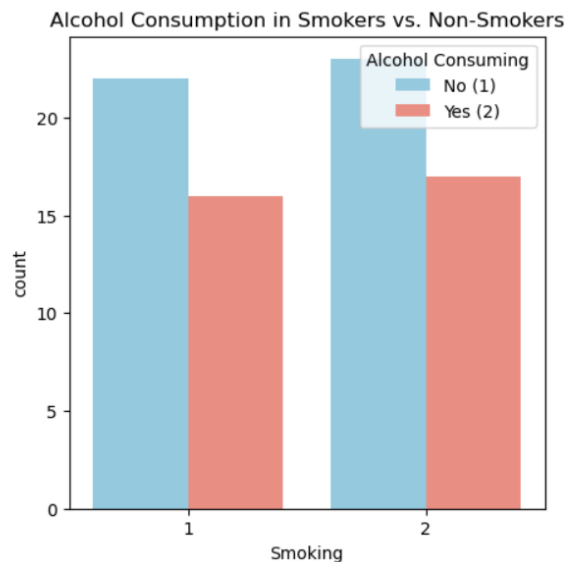
Smoking:

We can see that there are fewer cases of lung cancer in non-smokers against the larger number of lung cancer cases in the population of smokers. This co-insides with the notion that smoking as a carcinogen is predictive of lung cancer, however the difference is extremely small which is odd.

Alcohol Consuming:

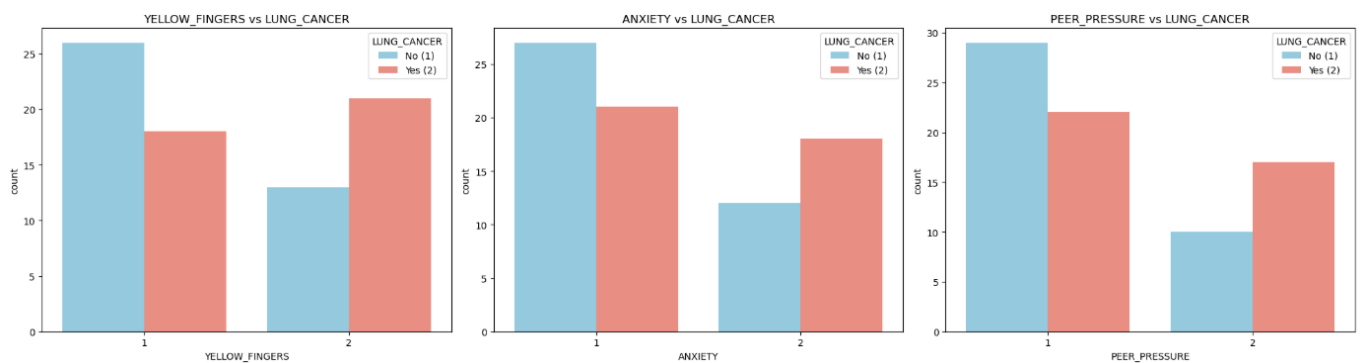
Far fewer non-alcohol consumers are diagnosed with lung cancer compared to the HUGE increase in lung cancer in alcohol consumers. This is expected as a carcinogen but rather odd that it is far more predictive than smoking.

Smoking against alcohol:



Alcohol consumption is spread equally against smokers and non-smokers.

Contributing factors:



Yellow fingers:

Less lung cancer in individuals with non-yellow fingers, and much more lung cancer in individuals with yellow fingers. First this is quite odd. Yellow fingers are an indication of smoking, so you'd think that the difference in cancer of smokers and those with yellow fingers would be similar yet here it is not. After some minor critical thinking it's clear to deduce that yellow fingers indicate long-term smoking which has proven to increase the risk of lung cancer. This means that most smokers in the dataset are short term smokers reducing the predictive potential.

Anxiety:

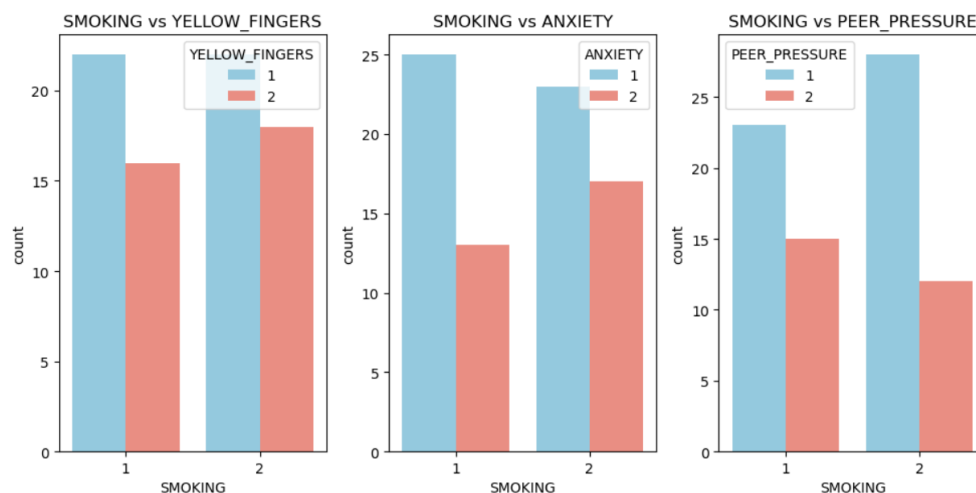
The disparity of lung cancer between those with and without lung cancer is similar to what's seen with yellow fingers. Meaning it's a strong predictor.

Peer pressure:

We see the exact same disparity as with the other carcinogen contributing factors. Peer pressure has predictive value.

Carcinogens on contributing factors:

Smoking-



Yellow fingers:

Somehow the distribution of yellow fingers among smokers and non-smokers is extremely similar. This could mean that a lot of smokers are new and have not yet stained their fingers, and non-smokers either have lied about their smoking status or quit smoking after long term. This can indicate why this feature has such little predictive power.

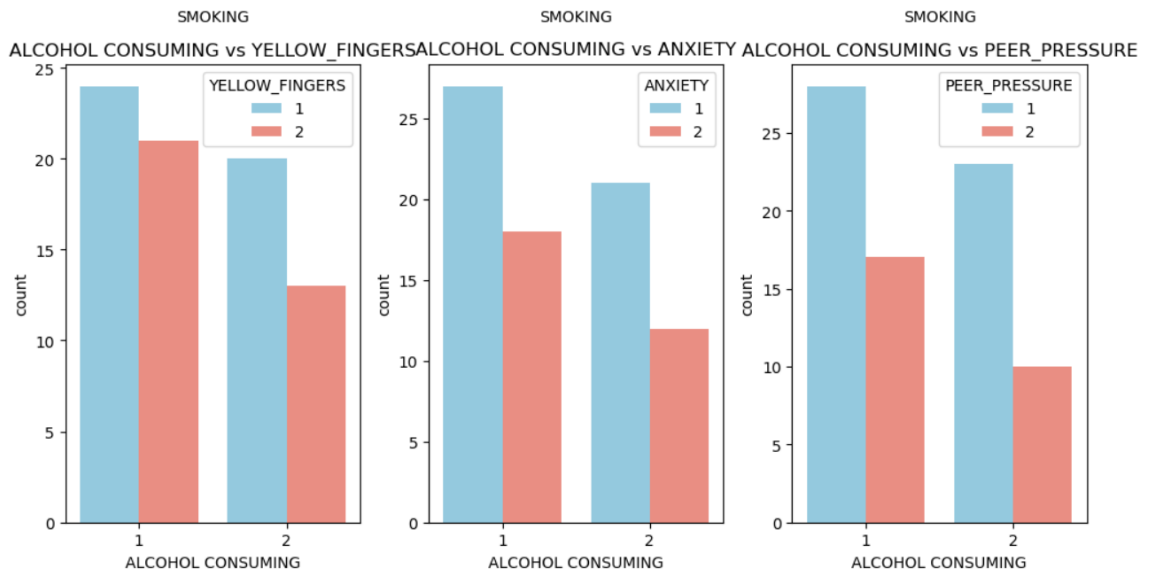
Anxiety:

Non-smokers seem to have less anxiety than smokers. This is just as expected.

Peer pressure:

Non-smokers experience more peer pressure than smokers. This makes sense since smokers can't be peer pressured into smoking since they already smoker.

Alcohol consumption -



Yellow fingers:

Yellow fingers are somehow more present in individuals who don't consume alcohol, which goes against the earlier notion that alcohol may indicate smoking as the two are commonly linked.

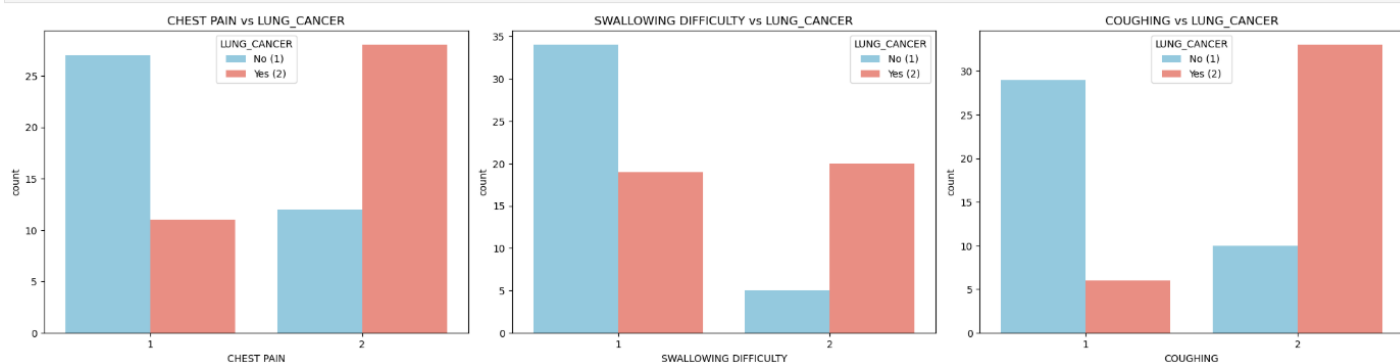
Anxiety:

Alcohol consumption has lower anxiety rates, which is the opposite of higher anxiety in smokers. Alcohol consumption seems to decrease anxiety.

Peer pressure:

Just like smokers, less peer pressure is expected in alcohol consumers as they already consume it, therefore there's nothing to be pressured into.

Symptoms



Chest Pain:

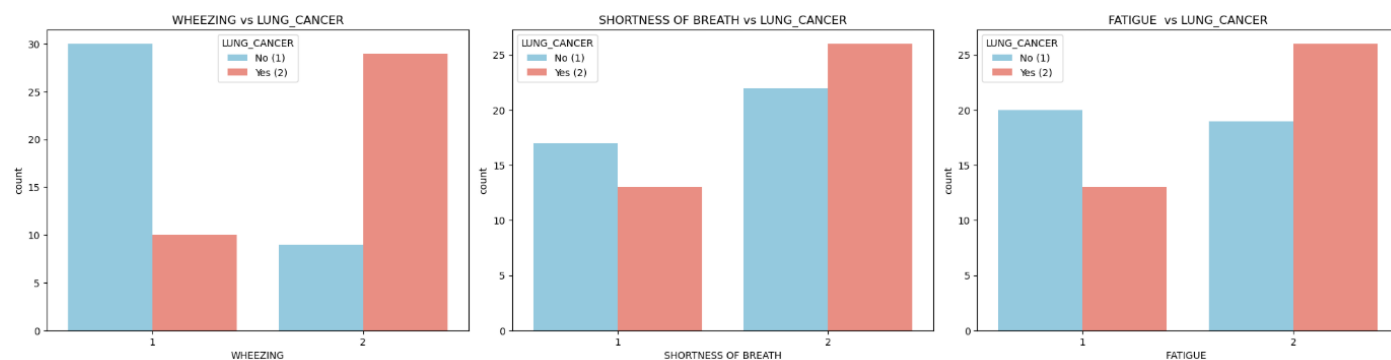
Individuals without chest pain have a lower number of lung cancer cases, as opposed to those with chest pain seem to have a huge spike in it. This is expected as it's a typical symptom of lung cancer.

Swallowing difficulty:

The same disparity can be seen here as in chest pain, however its noted that there are much less difficulty in swallowing, making it a rarer symptom and thus more likely to be a predictive element.

Coughing:

Again, the same disparity as the other symptoms can be seen here, however those without the coughing symptom seems to have a very low number of lung cancer cases, as opposed to those with coughing symptoms. Coughing is a very strong predictor of lung cancer.



Wheezing:

Very similar to the difference seen in coughing, however more individuals without wheezing seem to be affected by lung cancer.

Shortness of breath:

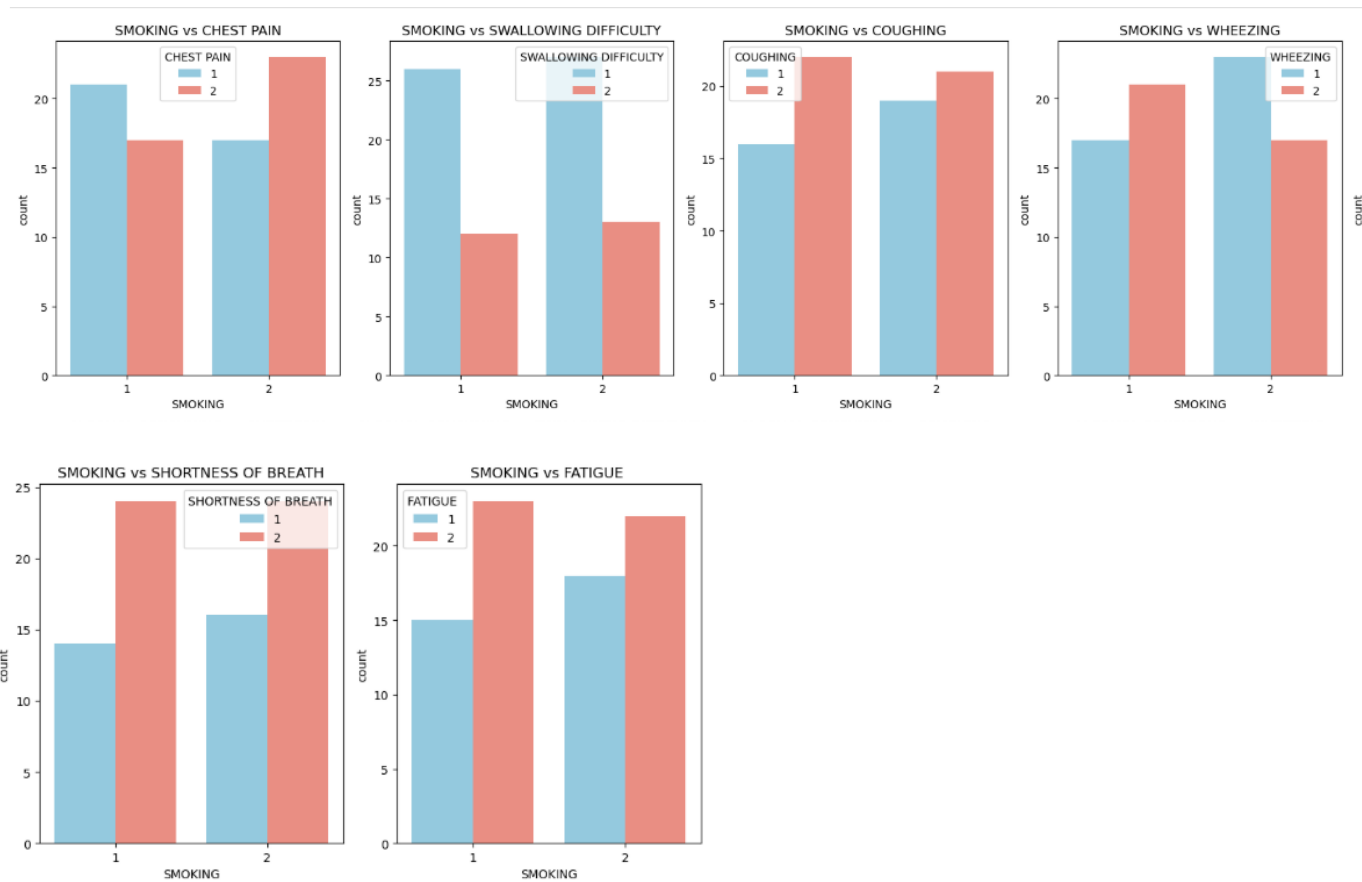
There is a much smaller difference between lung cancer in those without and without shortness of breathing. This means that there is very minor predictive value in this feature, despite it being a common symptom of lung cancer. This is unexpected but may be explained with further analysis.

Fatigue:

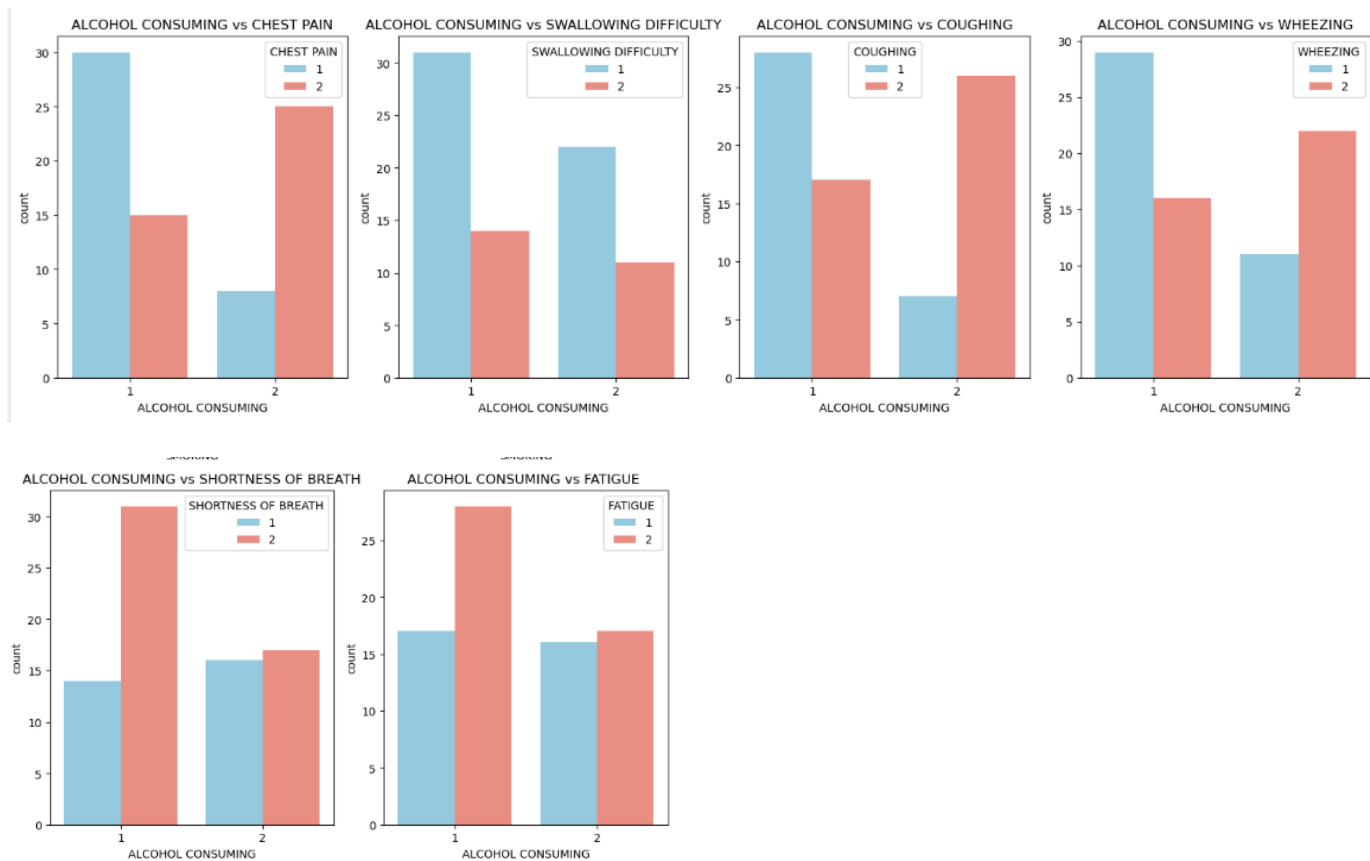
Very similar to the disparity seen in shortness of breath, and just as unexpected considering it's also a common symptom of lung cancer. This may also be explained soon in further analysis.

Symptoms against carcinogens:

Smoking:



Alcohol consuming:



Chest pain:

Smoking has a minor effect on chest pain, where nonsmokers have fewer cases of chest pain and smokers have more cases. Still this discrepancy is very small.

Strangely alcohol consumption has the same effect but with a much more significant difference, despite smoking being linked to lung health. Very strange.

Swallowing difficulty:

Both smoking and alcohol consumption have a similar effect on swallowing difficulty, neither seems to affect one another.

Coughing:

Regardless of smoking and non-smoking, both cases have a high number of coughing and even weirder, non-smokers seem to experience coughing more than smokers. Even more strange, alcohol consumption seems to have the effect you would expect from smoking, where those who consume alcohol have major increases in coughing as opposed to non-alcohol consumers.

Wheezing:

Similar effects to coughing with regards to smoker and alcohol consumption, however non-smokers seem to have higher cases of wheezing than smokers. Very strange.

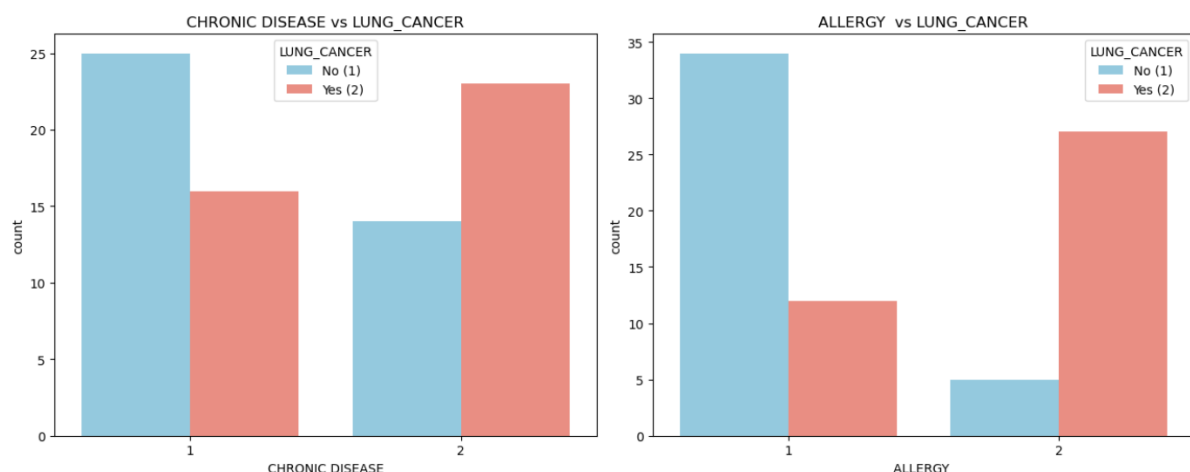
Shortness of breath:

Regardless of smoking or alcohol consumption cases have a high amount of shortness of breath.

Fatigue:

The same observation is made here as with shortness of breath.

Conditions:



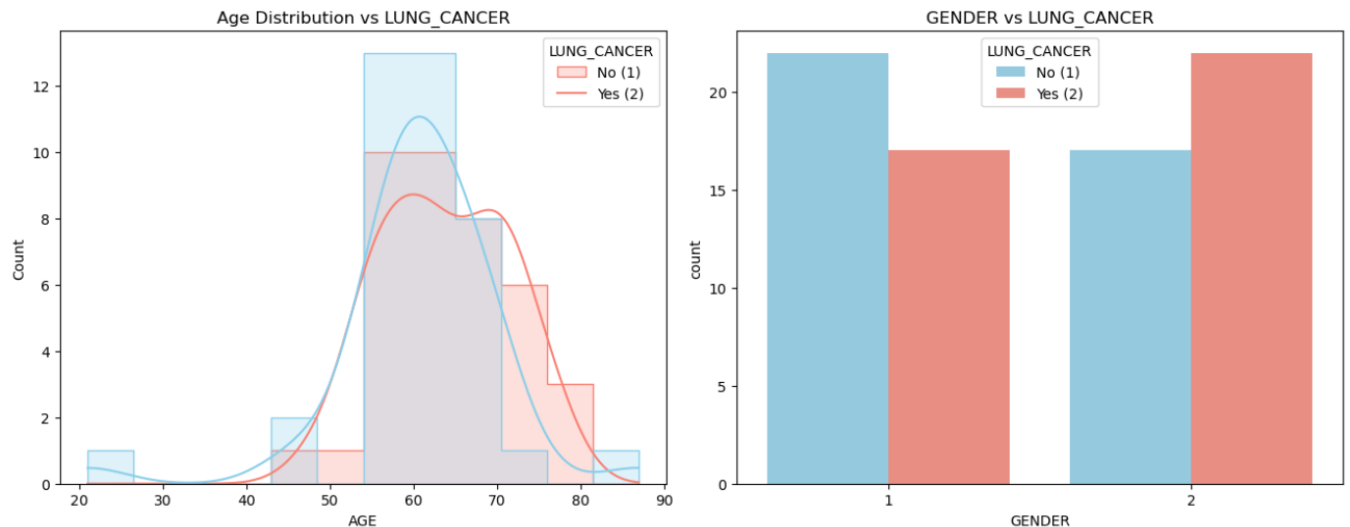
Chronic disease:

As expected, those with chronic diseases are likely to develop lung cancer.

Allergy:

Very strangely allergies have an even higher correlation with lung cancer. This deviates from the previous expectation that allergies aren't all that linked to lung cancer and may even reduce the chances.

Demographics:



Age:

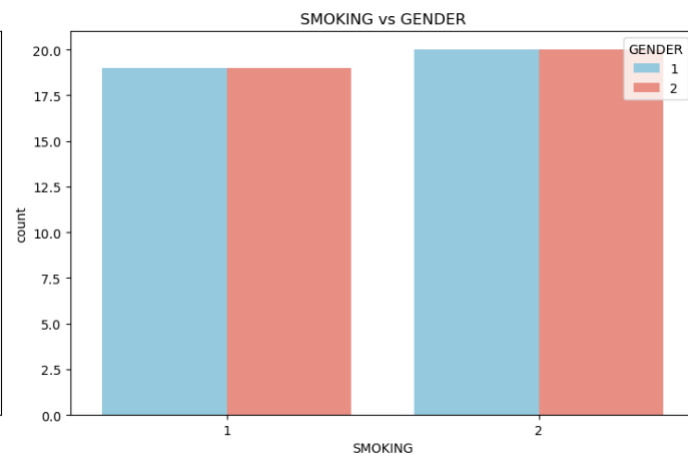
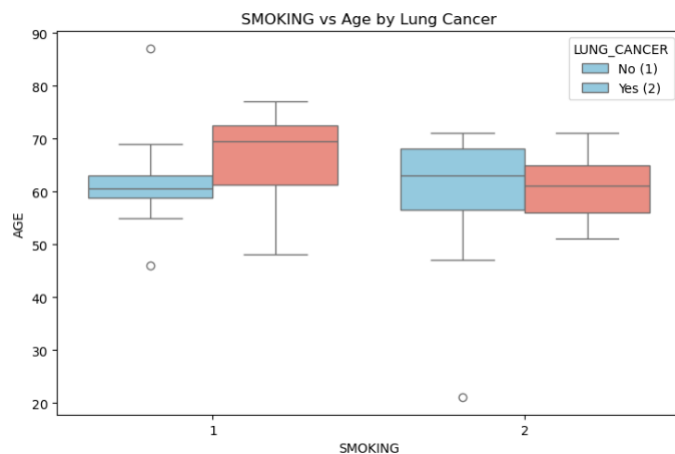
As age increases, we can see that lung cancer diagnoses also increase. This is expected.

Gender:

It seems that males (2) are more likely to be diagnosed with cancer than females. This observation has some predictive value.

Demographics against carcinogens:

Smoking-



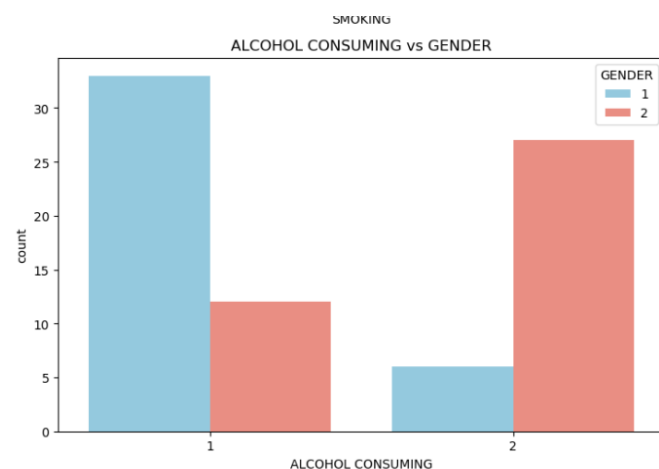
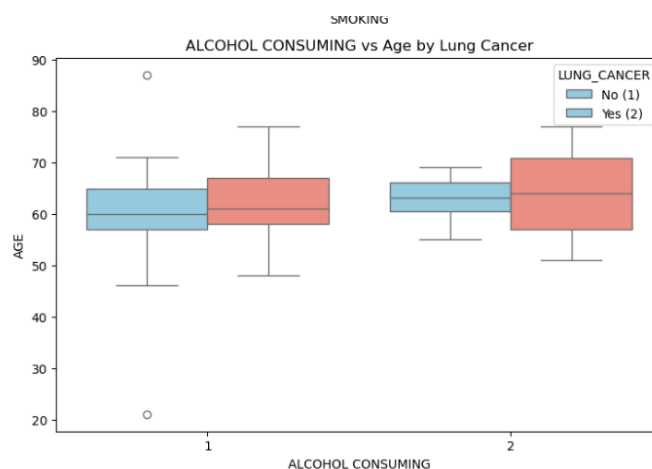
Age:

Non-smokers are generally older than smokers and seem to have a higher rate of lung cancer. It's odd that non-smokers have higher lung cancer cases, but it's also expected that the group with older individuals would have higher rates.

Gender:

Very minor differences in number of smokers and non-smokers for men and woman.

Alcohol consumption-



[]:

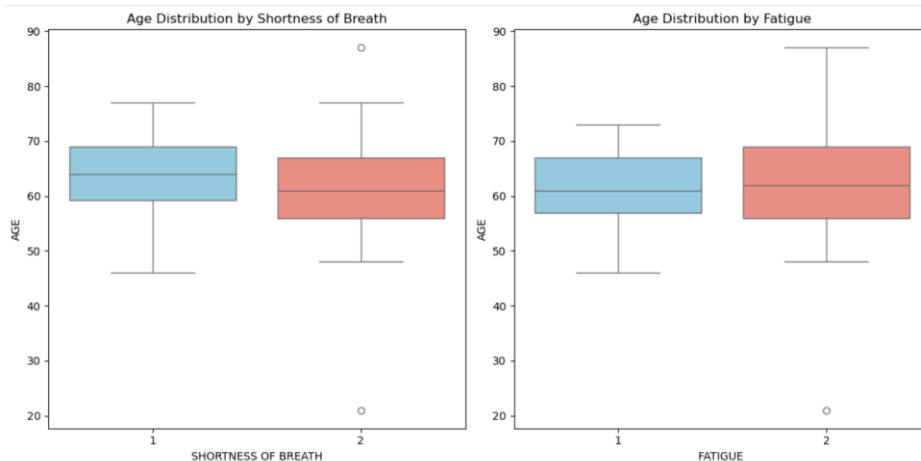
Age:

Non-alcohol consuming has a lower rate of lung cancer as well as a lower average age. This is completely expected based on previous findings.

Gender:

Males (2) consume significantly alcohol more than Females (1), comparing this to the graph that indicates more males have higher rates of lung cancer proves the expectation that alcohol consumption is a good indicator of lung cancer.

Age against Shortness of breath and fatigue:



Earlier in the EDA we noticed the odd indication that there is a high amount of fatigue and shortness of breath, with this plot we can see why. These symptoms are very common among individuals in the high age bracket which makes up almost the entire population. These symptoms are naturally occurring to people within this age group and therefore will have a high frequency.

Feature selection:

With our exploratory analysis complete, we've seen all possible trends and patterns to be found in the data. Now before we can train a model, we need to decide which features are to be included in the model to ensure the highest possible accuracy. The method used to perform feature selection will be using p values.

P values:

P values are a statistical value that indicates the likelihood of the observed data falling under the null hypothesis of a statistical test. In this case we're going to calculate the p values of each variable. The null hypothesis in this case would be if an independent variable affects the dependent variable which in this case is Lung_cancer. If the p value > 0.05 it's not a significant variable and we won't include it in our model (Banerjee Chandradip, 2023).

Using sklearn library, more specifically standard scaler we can build a small model and retrieve the p values of all the variables regarding how they affect the dependent variable.

These were the results of using the down sampled dataset:

```
Warning: Maximum number of iterations has been exceeded.  
Current function value: inf  
Iterations: 35
```

LinAlgError: Singular matrix

“Uh oh.”

Since the data was down sampled when trying to calculate P-values we have either one or two problems. Perfect multicollinearity, or there isn't enough variation in the dataset to create any predictions(Pontes, 2024).

To combat this problem, we calculate the p values based on the original skewed dataset, however we apply weights to the classes which output the following:

```
Optimization terminated successfully.
Current function value: 0.148720
Iterations 9

Logit Regression Results
=====
Dep. Variable:          LUNG_CANCER    No. Observations:          309
Model:                Logit          Df Residuals:              293
Method:                MLE           Df Model:                  15
Date:                  Thu, 29 May 2025    Pseudo R-squ.:            0.6077
Time:                  19:36:21          Log-Likelihood:           -45.955
converged:              True            LL-Null:                  -117.15
Covariance Type:        nonrobust        LLR p-value:              7.764e-23
=====
              coef    std err          z      P>|z|      [0.025    0.975]
-----
const         4.9926     0.765     6.523     0.000     3.492     6.493
x1             0.8809     0.348     2.530     0.011     0.199     1.563
x2             0.7004     0.397     1.765     0.078    -0.077     1.478
x3             0.6815     0.368     1.854     0.064    -0.039     1.402
x4             0.4439     0.406     1.092     0.275    -0.353     1.240
x5             0.8656     0.330     2.622     0.009     0.219     1.513
x6             0.2777     0.342     0.811     0.417    -0.393     0.949
x7             1.5581     0.564     2.763     0.006     0.453     2.663
x8             1.6347     0.529     3.090     0.002     0.598     2.672
x9             0.4800     0.414     1.158     0.247    -0.332     1.292
x10            -0.3497     0.365    -0.959     0.338    -1.064     0.365
x11            1.4402     0.387     3.721     0.000     0.682     2.199
x12            1.5957     0.444     3.593     0.000     0.725     2.466
x13            0.8178     0.382     2.141     0.032     0.069     1.566
x14            0.1788     0.278     0.643     0.520    -0.366     0.724
x15            -0.2627     0.354    -0.742     0.458    -0.957     0.431
=====

Possibly complete quasi-separation: A fraction 0.12 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.

Feature p-values:
x1      0.011396
x2      0.077616
x3      0.063777
x4      0.274672
x5      0.008740
x6      0.417213
x7      0.005722
x8      0.002003
x9      0.246737
x10     0.337551
x11     0.000199
x12     0.000327
x13     0.032292
x14     0.520477
x15     0.458050
dtype: float64
```

From these results we can see that our strongest features are, x1, x5, x7, x8, x11, x12, x13. These features represent, SMOKING, PEER_PRESSURE, SWALLOWING DIFFICULTY, COUGHING, FATIGUE, CHRONIC DISEASE, ALLERGY.

Some of these are expected such as coughing, difficulty swallowing, allergy, smoking and peer pressure. Unfortunately, some predictors such as alcohol consumption and yellow fingers are not strong enough predictors. This may be because of overlapping co-linearity discovered by the logistic model which ultimately reduced the predictive power of these features.

Model training:

With our features selected it's more or less time to start training models.

Firstly, we can't use the downsized model since it's far too small, and for this reason we're going to use smote.

SMOTE:

Smote is a technique used to help fix imbalances in datasets. It does this by creating artificial entries for the minority class. It's in the name, Synthetic Minority Oversampling Technique. This will allow us to get the most accurate model training possible. Of course, we'll only use smote on the training set of data (Blagus and Lusa, 2013).

Model accuracy metrics:

To determine the accuracy of each model we will be using classification reports from Sklearn as well confusion matrices.

Classification report:

This is a combination of different metrics that indicate how well a model has done. It consists of the following:

Precision:

the ratio of true positives against the number of predicted positives. Indicates how many of the actual positive cases the model correctly predicted (GeeksforGeeks, 2025).

Recall:

the ratio of true positives against the number of actual positives. It indicates how many true positives the model got right (GeeksforGeeks, 2025).

F1 Score:

The balanced mean of the precision and recall, this wholly represents the performance of the model (GeeksforGeeks, 2025).

Support:

The number of samples of a class within the test set (GeeksforGeeks, 2025).

Accuracy: A combination of all metrics to give an overall score for the models accuracy (GeeksforGeeks, 2025).

Confusion matrix:

A table that indicates the total numbers of correct and incorrect predictions made by the model. The table has four sections(GeeksforGeeks, 2025):

True positive: Correct true prediction.

True negative: Correct false prediction.

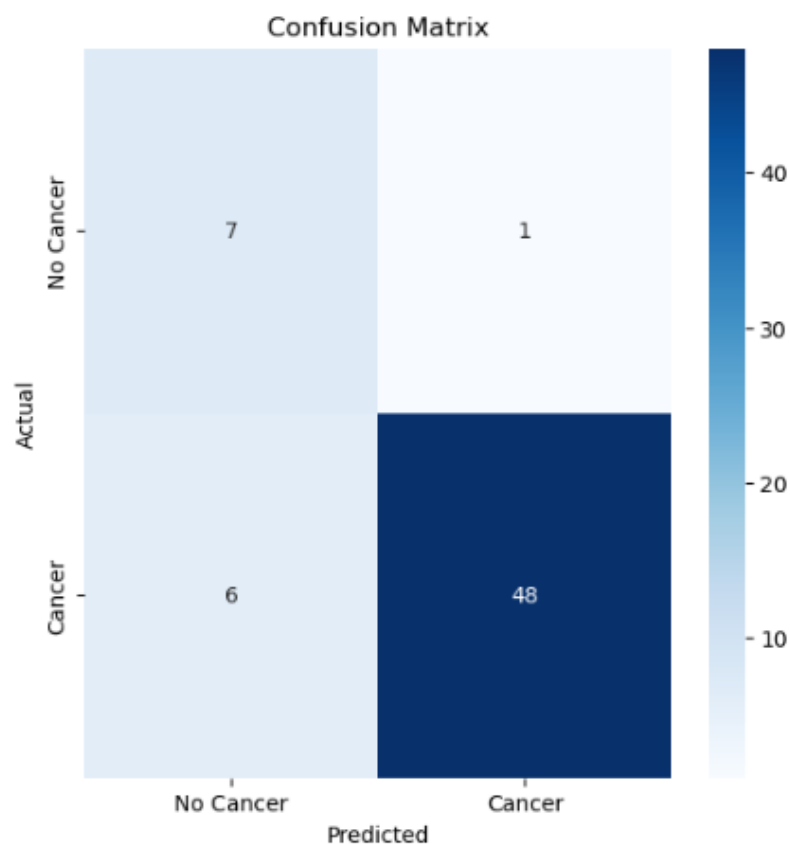
False positive: Incorrect true prediction.

False negative Incorrect false prediction (GeeksforGeeks, 2025).

Classification results:

Logistic regression:

| Classification Report: | | | | |
|------------------------|-----------|--------|----------|---------|
| | precision | recall | f1-score | support |
| 0 | 0.54 | 0.88 | 0.67 | 8 |
| 1 | 0.98 | 0.89 | 0.93 | 54 |
| accuracy | | | 0.89 | 62 |
| macro avg | 0.76 | 0.88 | 0.80 | 62 |
| weighted avg | 0.92 | 0.89 | 0.90 | 62 |



KNN:

(n=5)

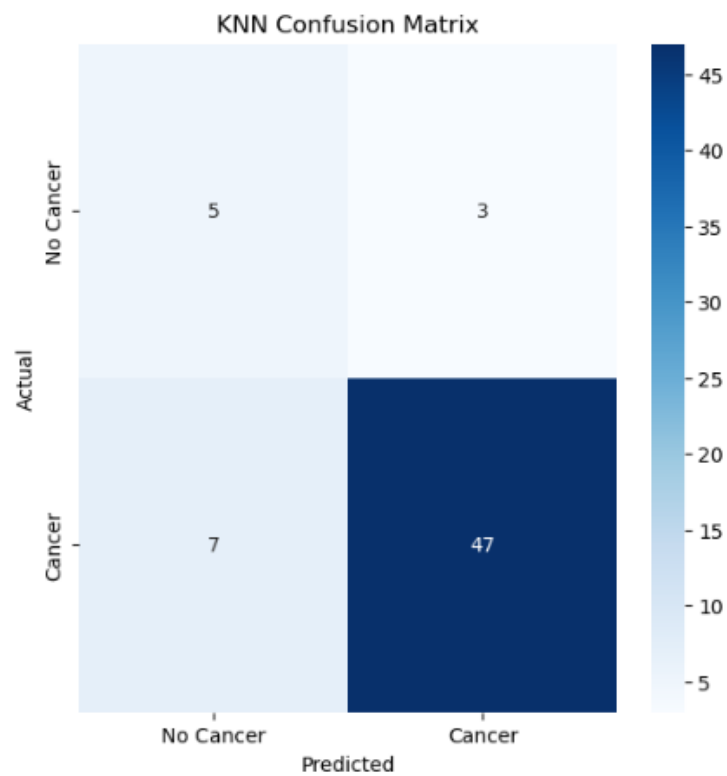
```

Classification Report:
      precision    recall  f1-score   support

     0       0.42      0.62      0.50         8
     1       0.94      0.87      0.90        54

 accuracy      0.84         62
 macro avg      0.68      0.75      0.70         62
 weighted avg      0.87      0.84      0.85         62

```



(n = 3)

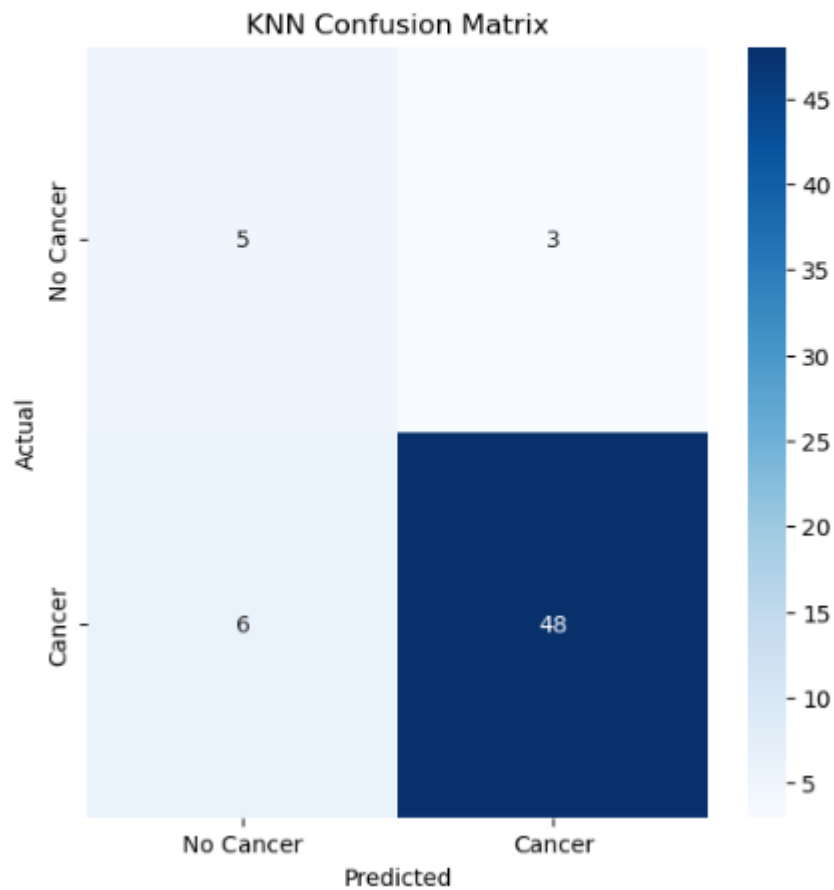
```

Classification Report:
              precision    recall  f1-score   support

     0       0.45        0.62        0.53         8
     1       0.94        0.89        0.91        54

 accuracy          0.85         62
 macro avg       0.70        0.76        0.72         62
 weighted avg    0.88        0.85        0.86         62

```



Decision tree:

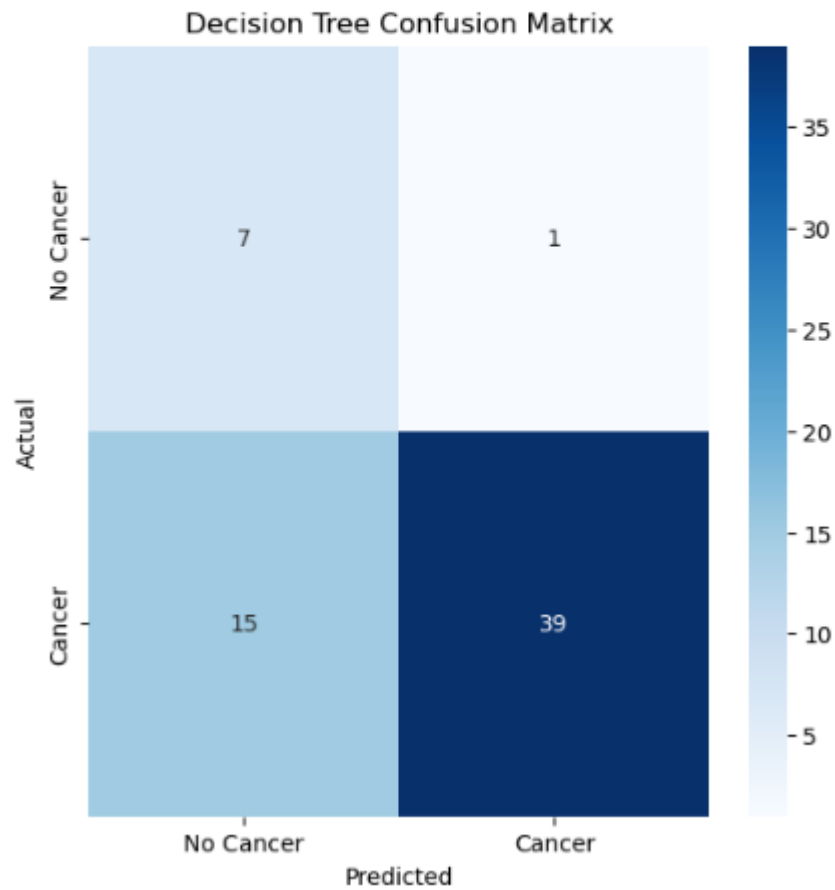
```

Classification Report:
              precision    recall  f1-score   support

     0       0.32         0.88      0.47         8
     1       0.97         0.72      0.83        54

 accuracy          0.74         62
 macro avg         0.65         0.80      0.65         62
 weighted avg      0.89         0.74      0.78         62

```



Random forest:

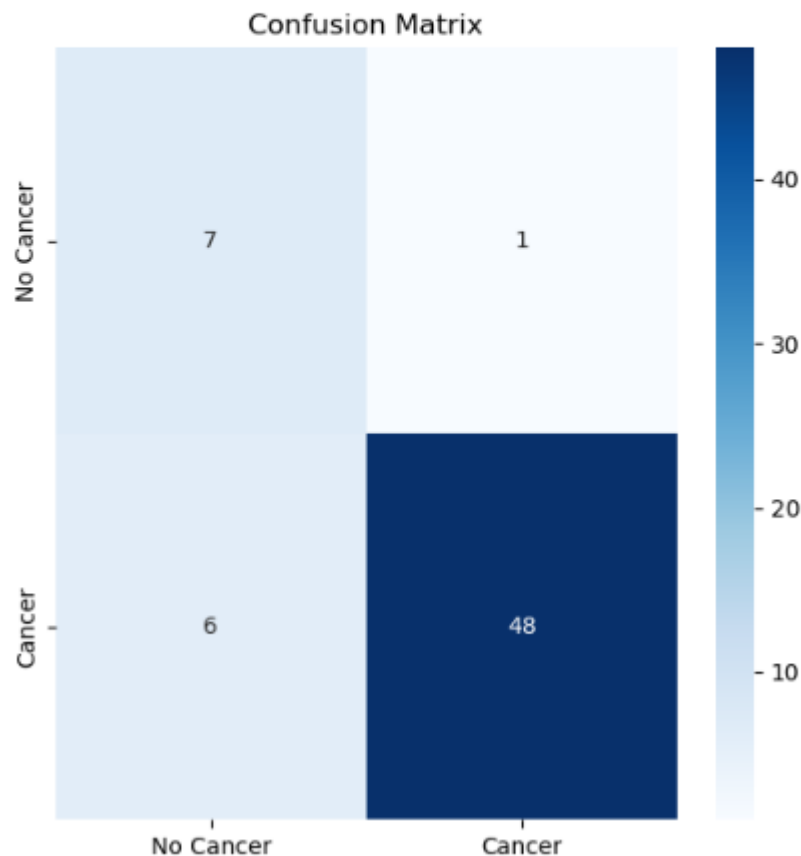
```

Classification Report:
              precision    recall  f1-score   support

     0       0.54         0.88     0.67         8
     1       0.98         0.89     0.93        54

 accuracy          0.89         62
 macro avg       0.76         0.88     0.80         62
 weighted avg    0.92         0.89     0.90         62

```



Conclusion:

The best trained classification model is the random forest model.

Regarding both the majority class (1), and the minority class (0), the model was able to achieve the highest F1 score for both classes. The model was also able to get the best all round accuracy score of 0.89.

References:

- Anon. 2024. *Univariate, Bivariate and Multivariate data and its analysis* | *GeeksforGeeks*. [online] Available at: <<https://www.geeksforgeeks.org/univariate-bivariate-and-multivariate-data-and-its-analysis/>> [Accessed 25 April 2025].
- Anon. 2025. *Pearson Correlation and Linear Regression*. [online] Available at: <<https://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/>> [Accessed 25 April 2025].
- Banerjee Chandradip, 2023. *P value and Feature Selection. P value and Feature Selection* | by Chandradip Banerjee | *Medium*. [online] Available at: <<https://medium.com/@chandradip93/p-value-and-feature-selection-629bec71d828>> [Accessed 25 April 2025].
- Bhandari, A., 2025. *Multicollinearity Explained: Causes, Effects & VIF Detection*. [online] Available at: <<https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/>> [Accessed 29 May 2025].
- Blagus, R. and Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, [online] 14. <https://doi.org/10.1186/1471-2105-14-106>.
- Cleveland Clinic, 2025. *Cancer Fatigue: What It Feels Like & How To Overcome It*. [online] Available at: <<https://my.clevelandclinic.org/health/diseases/5230-cancer-fatigue>> [Accessed 29 May 2025].
- encord, 2025. *Balanced and Imbalanced Datasets in Machine Learning [Full Introduction]*. [online] Available at: <<https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/>> [Accessed 29 May 2025].
- GeeksforGeeks, 2025. *Compute Classification Report and Confusion Matrix in Python* | *GeeksforGeeks*. [online] Available at: <<https://www.geeksforgeeks.org/compute-classification-report-and-confusion-matrix-in-python/>> [Accessed 29 May 2025].
- healthline, 2025. *Lung Cancer Symptoms: Coughing, Wheezing and More*. [online] Available at: <<https://www.healthline.com/health/lung-cancer-symptoms>> [Accessed 29 May 2025].
- Khan, Q.U., Rehman, M.U., Abbasi, M.A.A., Shiekh, R.R., Nazir, M., Raja, S.K., Akbar, A., Tasneem, S., Jadoon, S.K. and Alvi, S., 2024. Correlation between allergic rhinitis or hay fever and lung cancer: A systematic review and meta-analysis. *Medicine*, [online] 103(20), p.e38197. <https://doi.org/10.1097/MD.00000000000038197>.
- Lee, G., Walser, T.C. and Dubinett, S.M., 2009. Chronic inflammation, chronic obstructive pulmonary disease, and lung cancer. *Current Opinion in Pulmonary*

Medicine, [online] 15(4), pp.303–307.

<https://doi.org/10.1097/MCP.0B013E32832C975A>,.

Leshargie, C.T., Alebel, A., Kibret, G.D., Birhanu, M.Y., Mulugeta, H., Malloy, P., Wagnaw, F., Ewunetie, A.A., Ketema, D.B., Aderaw, A., Assemie, M.A., Kassa, G.M., Petrucka, P. and Arora, A., 2019. The impact of peer pressure on cigarette smoking among high school and university students in Ethiopia: A systemic review and meta-analysis. *PLoS ONE*, [online] 14(10), p.e0222572.

<https://doi.org/10.1371/JOURNAL.PONE.0222572>.

National Cancer Institute, 2025. *Alcohol and Cancer Risk Fact Sheet - NCI*. [online] Available at: <<https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet>> [Accessed 29 May 2025].

NHS, 2025. *Stopping smoking for your mental health - NHS*. [online] Available at: <<https://www.nhs.uk/live-well/quit-smoking/stopping-smoking-mental-health-benefits/>> [Accessed 29 May 2025].

Northrup, T.F., Stotts, A.L., Suchting, R., Khan, A.M., Klawans, M.R., Green, C., Hoh, E., Hovell, M.F., Matt, G.E. and Quintana, P.J.E., 2022. Handwashing Results in Incomplete Nicotine Removal from Fingers of Individuals who Smoke: A Randomized Controlled Experiment. *American Journal of Perinatology*, 39(15), pp.1634–1642.

<https://doi.org/10.1055/S-0041-1736287>.

Pontes, L., 2024. *Errors with Numpy and Scipy about Matrix Conceptions | by Luciano Pontes | Medium*. [online] Available at: <<https://medium.com/@lutcho/errors-with-numpy-and-scipy-about-matrix-conceptions-56c8bc1d2e56>> [Accessed 29 May 2025].

Walser, T., Cui, X., Yanagawa, J., Lee, J.M., Heinrich, E., Lee, G., Sharma, S. and Dubinett, S.M., 2008. Smoking and Lung Cancer: The Role of Inflammation. *Proceedings of the American Thoracic Society*, [online] 5(8), p.811.

<https://doi.org/10.1513/PATS.200809-100TH>.