

Analysis of Lung Cancer Factors to Predict Risk

Jonathan Liu¹, Sameer Mohammad¹, Bhargav Varidi¹, Teja Pavani Jyesta¹, Lavanya Ranganatham¹, and Pallavi Vandanapu¹

¹Indiana University-Purdue University, Indianapolis, IN 46202, USA

jonwliu@iu.edu, sammoha@iu.edu, bvaridi@iu.edu, tjyesta@iu.edu, laranga@iu.edu, pvandana@iu.edu

Abstract. This project investigates the influence of multiple lung cancer risk factors. The goal is to identify the factors that most strongly predict lung cancer severity. In data analysis, we determined the associations between factors and lung cancer severity to reveal the strength of influence of each factor. We also created several predictive models to evaluate lung cancer risk of new patients based on patterns of our dataset. Overall, our project strives to increase understanding of major lung cancer risks to encourage early action and facilitate preventative medicine.

Keywords. Lung cancer, risk factors, severity, prevention, data cleaning, data analysis, data visualization, Python, Seaborn

1 Project Scope

1.1 Introduction

According to the American Cancer Society (2023), cancer is the second leading cause of death in the United States, and lung cancer is one of the deadliest forms. Even during early localized stages, non-small cell and small cell lung cancers only have 65% and 30% 5-year survival rates, respectively. These statistics suggest the ineffectiveness of treatment during later stages of disease development. Once lung cancer is detected, the adverse health impacts may already be irreversible. As a result, healthcare and research should focus on detecting key risk factors early and educating populations to encourage behavioral changes that mitigate such risk factors.

The ACS (2023) notes that the primary lung cancer risk is smoking. They also list secondhand smoke, environmental pollution, and carcinogen exposure as other major contributors. Other sources corroborate these findings, as Malhotra et al. (2016), indicates health behaviors, environment, and genetics as the largest contributing factors to lung cancer. Based on previous research, we predict that smoking, passive smoking, air pollution, dust allergy, occupational hazards, and genetic risk will have the highest association with lung cancer development. Other factors such as obesity, fatigue, and wheezing will have lower associations with lung cancer development.

1.2 Aim

The goal of our project is to evaluate major risk factors that correlate with lung cancer severity and construct predictive models to guide preventative medicine. Our project has three aims:

- To determine whether any risk factors are associated with lung cancer severity.
- To identify interactions between risk factors.
- To construct models that predict lung cancer risk.

Our hypotheses are as follows:

- Null Hypothesis - None of the examined risk factors are associated with lung cancer development, so these factors cannot be used to predict lung cancer risk in new patients.
- Alternative Hypothesis - One or more of the examined risk factors, mainly smoking, air pollution, and genetics, are associated with lung cancer development, so these factors can be used to predict lung cancer risk in new patients.

1.3 Purpose

The primary purpose of this study is to conduct a detailed analysis of potential risk factors for lung cancer to quantify the strength of their associations with the development and severity of disease. The dataset of 1,000 lung cancer patients includes information about smoking history, occupational hazards, genetics, and comorbidities. Statistical modeling will be used to identify the greatest risks. Beyond simply estimating the correlation of individual factors, we will also investigate interaction effects between risk factors that may act synergistically to amplify risk. For example, combining smoking with air pollution may contribute compounded lung cancer risk compared to each factor acting alone. A thorough understanding of risk relationships will provide the foundation for an accurate multifactorial model to predict an individual's overall lung cancer risk based on their unique risk factor profile. Identifying the most influential modifiable factors can guide public health initiatives to target efforts for reducing lung cancer incidence through primary prevention.

2 Methodology

2.1 Project Steps

This project focuses on comparing risk factors such as air pollution, genetic risk, and smoking with the severity of lung cancer by using various Python tools. Factor evaluation is then used to construct predictive models for lung cancer risk. Primary processes include:

1. Data Collection
2. Data Processing
3. Data Analysis and Visualization
4. Data Modeling
5. Results and Interpretation

2.2 Original Responsibilities

Below are the original task assignments from our project proposal earlier in the semester.

Name	Roles
Jonathan Liu	Organize meeting times, host zoom calls, propose initial ideas, coordinate discussion, write drafts, and submit group assignments
Teja Pavani Jyesta	Analyze the dataset, choose appropriate data analysis techniques and identify relevant variables and develop predictive models for disease prognosis or diagnosis.
Sameer Mohammad	Provide clinical insights, interpret medical records, assist in defining the research questions and reports in the dataset.
Bhargav Varidi	Draft and edit the proposal document. Summarize the project objectives, methodology and expected outcomes. Create clear and concise documentation for methods and results. Ensure proper formatting and citation of references.
Lavanya Ranganatham	Assist in designing study, including sample size calculation. Ensure validity, reliability, and interpret statistical results in a meaningful way.

Pallavi Vandanapu	Draft and edit the proposal document. Summarize the project objectives, methodology and expected outcomes. Create clear and concise documentation for methods and results. Ensure proper formatting and citation of references.
-------------------	---

Fig. 1. Originally anticipated project responsibilities for team members

2.3 Actual Contributions

While we initially attempted to distribute tasks evenly, we did not understand the concepts necessary to complete our project at the time of project proposal submission. Class lectures regarding data processing, analysis, modeling, and visualization all occurred after the proposal was due. Once we understood the knowledge and procedures necessary for project tasks, we were better able to delegate responsibilities. Below are the actual contributions of each member:

Name	Roles
Jonathan Liu	Project Manager, Organize/Lead meetings, Create project documents, Construct presentation slides/scripts/timing, Proofread/Revise code (Data Cleaning, Data Analysis, Data Visualization, and Data Modeling), Write final proposal (Abstract, Introduction, Data Cleaning, Feature Representation, Modeling, Results, Discussion, References), Proofread other sections.
Sameer Mohammad	Provide clinical insights, interpret medical records, assist in defining the research questions and reports in the dataset, Draft and edit the proposal document, Co-lead coding (Data Cleaning, Data Analysis, Data Visualization, and Data Modeling), Assist in designing study, including sample size calculation. Ensure validity, reliability, and interpret statistical results in a meaningful way. Create clear and concise documentation for methods and results. Ensure proper formatting. Write and read the final proposal.
Bhargav Varidi	Provide clinical insights, interpret medical records, assist in defining the research questions and reports in the dataset, Draft and edit the proposal document, Co-lead coding (Data Cleaning, Data Analysis, Data Visualization, and Data Modeling), write and read the final proposal.
Teja Pavani Jyesta	Assigned distinct responsibilities to team members based on their original roles in the project proposal, Drafted and edited the proposal document, Performed few machine learning models for data modeling and made comparisons from one model to other, Revised and made necessary changes in the code (Data Modeling)
Lavanya Ranganatham	Provide an appropriate study design and use random sampling for representativeness. Calculated the sample size considering confidence level, effect size, and power. Ensured precise measurement of variables, established reliability through testing, and validated through expert input.
Pallavi Vandanapu	Draft and edit the proposal document, project presentation

Fig. 2. Actual project contributions of team members

2.4 Project Challenges

- Data completeness and quality: inaccurate, missing, or incomplete patient data might have problems that make analysis less reliable. There is a need for extensive data cleansing and validation.

- Intricate modeling using a large number of factors: Overfitting predictive models is a danger when there are many risk factors or variables. Even with numerous input parameters, model interpretability degrades and regularization techniques are necessary.
- Subjectivity in grading of severity: The classification of cancer severity as low, medium, or high depends on clinical judgements that may be arbitrary at the time of data collection. It would be necessary to use standard criteria.
- Changing demographics: Over time, there may be changes in the population's distribution of risk variables. Without regular re-validation, the predictive model runs the danger of having decreased generalizability.
- Ethical considerations: An ethical review would be necessary for this study because it uses retrospective patient data without consent. Precautions for data security and privacy are crucial.
- Resource limitations: With just six team members and a three-month timetable, it would not be possible to finish a thorough investigation and software development without an adequate computer infrastructure.

3 Data Collection

This retrospective data analysis comprises a thorough investigation of an already-existing dataset that contains 1,000 lung cancer diagnoses. The dataset was obtained from the Data.World Lung cancer dataset, which was assembled by Prithivraj in 2017. This dataset includes 23 unique parameters that are regarded as independent variables, all of which may have an impact on the prognosis and course of lung cancer. These variables may include a range of items, including lifestyle decisions, treatment options, tumor features, and other medical data. Understanding the relationship between these independent variables and the disease's severity level, which serves as the dependent variable, is the main goal of this analysis. Examining this large dataset will provide important insights that may help find important predictors or correlations related to the severity of lung cancer. This will help understand how the disease progresses and open up new possibilities for better patient care protocols, treatment approaches, or diagnostics.

4 Data Processing

4.1 Data Extraction

The Kaggle data science platform provided the dataset of patients with lung cancer that was examined in this study. To facilitate study, the "Lung Cancer Data" dataset was made freely available under a Creative Commons license after being released by user Prithivraj (2017). With the use of this pre-existing dataset, a library of 1,000 lung cancer patients' clinical symptoms, comorbidities, environments, lifestyle choices, genetics, and other possible risk factors are available. More sophisticated analytics than are usually possible with single-center institutional data are made possible by using Kaggle to access an external big data source.

Our study was able to analyze a large lung cancer cohort orders of magnitude larger than most comparable investigations in the literature by utilizing this platform of crowdsourced medical data. Relying just on patient records that have been submitted to Kaggle, however, presents additional issues with data confidentiality, completeness, standardization, and quality that call for mitigating measures. However, the abundance of variables for each patient allows for more detailed multivariate analysis into intricate correlations between exposures and results. More trustworthy machine learning algorithms are trained for individualized risk assessments thanks to the amount of data. Open-access big data may be a very useful accelerator for repeatable predictive analytics when combined with appropriate data governance procedures.

4.2 Feature Representation

For each factor in our data, we evaluated scientific literature to confirm relevance to lung cancer. This way, we could potentially remove irrelevant data from analysis that have no predictive value and would only introduce irrelevant noise.

Firstly, American Cancer Society (2023) indicates that lung cancer risk is increased by negative health behaviors such as smoking, alcohol use, and poor diet. Also, existing health conditions such as genetic risk, dust

allergy, chronic lung disease, and obesity contribute risk. Thirdly, poor environmental conditions such as air pollution, occupational hazards, and passive smoking increase risk. Fourthly, common lung cancer symptoms include chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, dry cough, frequent cold, and snoring. Cleveland Clinic (2023) explains that fingernail clubbing can indicate poor oxygen levels and underlying lung disease. Finally, May et al. (2023) indicates that demographics such as old age and male gender face higher risk of lung cancer. In conclusion, based on existing scientific knowledge, every factor within our dataset is associated with lung cancer risk. As such, no factors require omission from analysis.

Based on our review, factors from our dataset fall into five categories. Demographics, behaviors, health conditions, environmental conditions, health conditions, and health symptoms.

Category	Factor
Demographics	Age, Gender
Behaviors	Alcohol Use, Balanced Diet, Smoking
Health Conditions	Dust Allergy, Genetic Risk, Chronic Lung Disease, Obesity
Environmental Conditions	Air Pollution, Occupational Hazards, Passive Smoking
Health Symptoms	Chest Pain, Coughing, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Fingernails, Frequent Cold, Dry Cough, Snoring

Fig. 3. Categorized lung cancer risk factors

4.3 Data Cleaning

Patient ID data was not expected to have associations with lung cancer risk and was dropped from analysis. Additionally, we noted that our gender data did not have labels available to indicate which values correlated with male and female. As a result, we also opted to remove gender from most analysis, visualization, and modeling, because any results involving gender data would have ambiguous meaning.

Data was cleaned by evaluating for errors such as duplicate patients or missing data values. Additionally, we noted a scaling discrepancy. Our lung cancer factors, the independent variables, utilized numerical scoring. Most factors utilized a 1-to-9 scale. However, our primary dependent variable, level of lung cancer severity, utilized a categorical low-medium-high scale. To establish consistent scales between our independent and dependent variables, we converted the categorical scale into a numerical scale to facilitate subsequent steps such as regression modeling.

```
# To identify the null values
df.isnull().sum()

# To identify the duplicate entries
df = df.drop_duplicates()

# Dropping the unwanted columns
df.drop(columns = ['Patient Id'], axis = 1, inplace=True)

#Converting categorical value into numerical values
df=df.replace({'Level':{'Low': 1, 'Medium': 5, 'High': 9}})
```

Fig. 4. Code for data cleaning

4.3 Data Import

In the process of importing data from the "cancer_patient_datasets.csv" file into a Jupyter Notebook (jpynb) for comprehensive data management, analysis, and visualization, several key steps were executed. Initially, the Pandas

library was employed to read the CSV file, loading the cancer patient datasets into a Pandas DataFrame within the Jupyter environment. This facilitated a structured representation of the data, allowing for efficient manipulation and exploration. Subsequently, data cleaning procedures were implemented to address missing values, outliers, and inconsistencies, ensuring the dataset's integrity.

Following the data cleaning phase, Pandas functionalities were leveraged for exploratory data analysis (EDA), allowing for statistical summaries, distribution analyses, and correlation assessments to unveil meaningful insights. Matplotlib and Seaborn were utilized for data visualization, generating informative plots and charts to visually communicate trends and patterns within the cancer patient datasets. The Jupyter Notebook also witnessed the application of scikit-learn or other relevant machine learning libraries for constructing data models, enabling predictive analyses and furthering our understanding of the underlying patterns in the cancer patient data. This holistic approach, encompassing data cleaning, analysis, visualization, and modeling, served to enhance the overall comprehension and utilization of the cancer patient datasets within the Jupyter Notebook environment.

4.4 Data Storage

Maintaining the lung cancer dataset in an organized format, such as CSV, implementing backup plans for data preservation, and implementing version control for tracking changes are essential for efficient data storage and administration. Encryption techniques and access restrictions should protect patient data with HIPAA and GDPR standards. Understanding and executing analyses is made easier with thorough descriptive documentation of the dataset's history, variables, and cleaning techniques. Sustainable storage options should also provide for future growth. Data protection should have a disaster recovery strategy in place, as well as procedures for safe disposal and data retention that conform with legal requirements. Effective analysis and insights derivation are facilitated by collaboration, clear documentation, frequent monitoring, and maintenance procedures that ensure the dataset's integrity, accessibility, and compliance throughout its lifecycle.

5 Data Analysis and Visualization

Python was used for data analysis, and both Python and Seaborn were used for visualization. Key Python packages include pandas, matplotlib, numpy, and scipy. Below are key steps we took to analyze and visualize our data.

- Shapiro-Wilk normality test
- Factor versus severity level correlations
- Factor versus factor correlations
- Spearman coefficient for correlation significance
- Visualizations (bar chart, heatmap, histogram, pie chart, forest plot, etc) to demonstrate data patterns

5.1 Normality

```
[ ] from scipy.stats import shapiro

factors = ['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']

for factor in factors:
    stat,p_value = shapiro(df[factor])
    print(factor)
    print("Shapiro-Wilk Stat =",stat,"and P-value = ",p_value)
```

Fig. 5. Code for shapiro-wilk normality test

To evaluate the distribution of each factor within our data, we used the shapiro-wilk normality test. The P-value for all 23 factors (independent variables) and severity (independent variable) was smaller than 0.05, indicating sufficient evidence to reject the null hypothesis that our data is normally distributed. In other words, our data does not follow normal distribution.

5.2 CORRELATION LUNG CANCER SEVERITY FACTOR

```
CORRELATION LUNG CANCER SEVERITY FOR EACH FACTOR

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Extracting the relevant columns for the analysis
factors = ['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']

subset_df = df[factors]

if subset_df.isnull().values.any():
    subset_df = subset_df.dropna()

correlation_with_severity = subset_df.corr()['Level']

correlation_with_severity = correlation_with_severity.sort_values(ascending=False)

print(correlation_with_severity)

# Visualize the correlations using a bar chart
plt.figure(figsize=(12, 6))
sns.barplot(x=correlation_with_severity.index, y=correlation_with_severity.values)
plt.xticks(rotation=45, ha='right')
plt.title('Correlation with Lung Cancer Severity for Each Factor')
plt.xlabel('Factors')
plt.ylabel('Correlation')
plt.show()
```

Fig. 6. Code for correlation lung cancer severity for each factor.

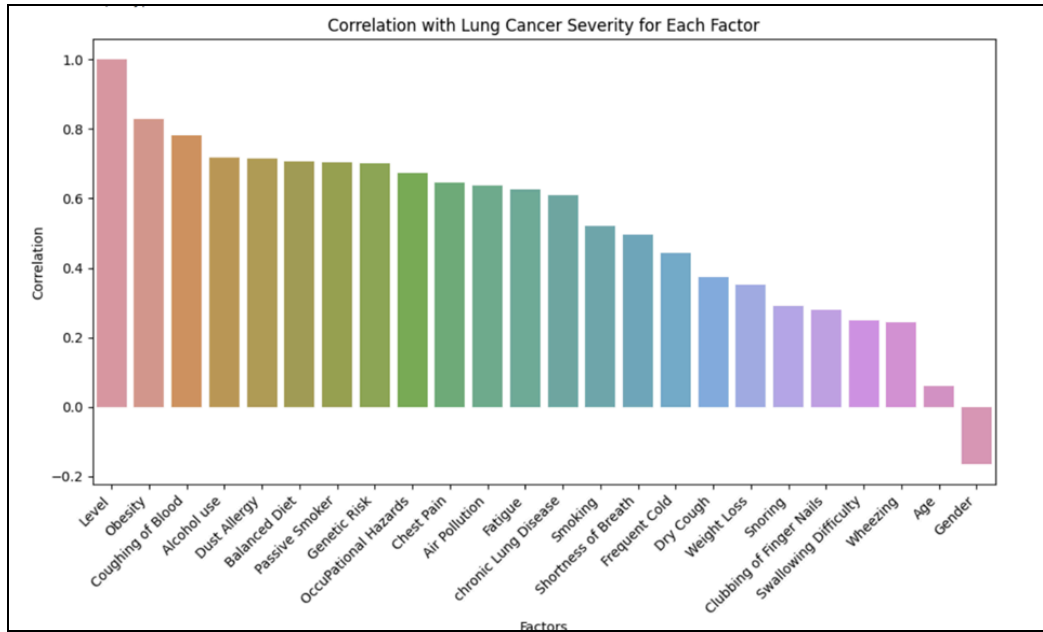


Fig. 7. Bar Graph showing Correlation with lung cancer Severity for each factor.

The bar graph displays the correlation between lung cancer severity and several factors. The correlation coefficient is a measure of the strength and direction of the connection between two variables. It ranges from -1 to 1, where a correlation of 1 indicates a perfect positive correlation, a correlation of -1 indicates a perfect negative correlation, and a correlation of 0 indicates no correlation.

Firstly, nearly all risk factors were found to have positive correlations with severity level. Interpreting our gender data is difficult due to the lack of labels. Considering standard correlation strengths cutoffs (weak <0.4, moderate 0.4-0.6, and strong >0.6) and our 5 categories of factors, we found that behaviors (alcohol use, balanced diet, smoking), environmental conditions (occupational hazards, air pollution), health conditions (obesity, dust allergy), and major health symptoms (coughing of blood, chest pain) to show stronger correlations. In contrast, demographics (age, gender) and mild symptoms (dry cough, snoring, wheezing) showed weaker correlations.

Based on scientific literature, we did expect positive correlations between these factors and severity level. However, the order of correlation strength was surprising. For example, smoking was ranked 13th, which was much lower than expected. These patterns are further explored in our discussion section below.

Also, we suspect that “balanced diet” was mislabelled in our dataset, because having a good diet would be expected to have a negative correlation with severity. However, our data shows that a “balanced diet” is positively correlated with severity level.

It is important to note that correlation alone does not imply causation. Two correlated factors do not necessarily cause each other. However, the strong correlation between lung cancer severity and smoking, occupational hazards, and air pollution suggests that these factors play a significant role in the development and progression of lung cancer. It is also important to consider that the bar graph is based on a single study, and the results may not be applicable to all populations.

Statistical Significance of Factor versus Severity Correlations

```
from scipy.stats import spearmanr
for factor in factors:
    correlation_coefficient, p_value = spearmanr(df[factor],df['Level'])
    print(factor)
    print('Spearman Correlation Coefficient =',correlation_coefficient,"and P-value = ",p_value)
```

Fig. 8. Code for Spearman correlation between each factor and severity level

In addition to correlation strengths explored above, we also calculated the statistical significance of the correlations between factors and severity level. Based on our normality test that strongly suggests our data is not normally distributed, we opted to measure correlation significance with the Spearman correlation coefficient. We found all P-values to be much smaller than 0.05, suggesting sufficient evidence to reject the null hypothesis and conclude that statistically significant correlations exist between factors and severity level.

5.3 CORRELATION MATRIX HEATMAP

CORRELATION MATRIX HEATMAP

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

# Extracting the relevant columns for the analysis
factors = ['Age', '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']

correlation_matrix = df[factors].corr()

# Visualize the correlation matrix heatmap
plt.figure(figsize=(20, 9))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix Heatmap')
plt.show()
```

Fig. 9. Code for Correlation matrix Heatmap

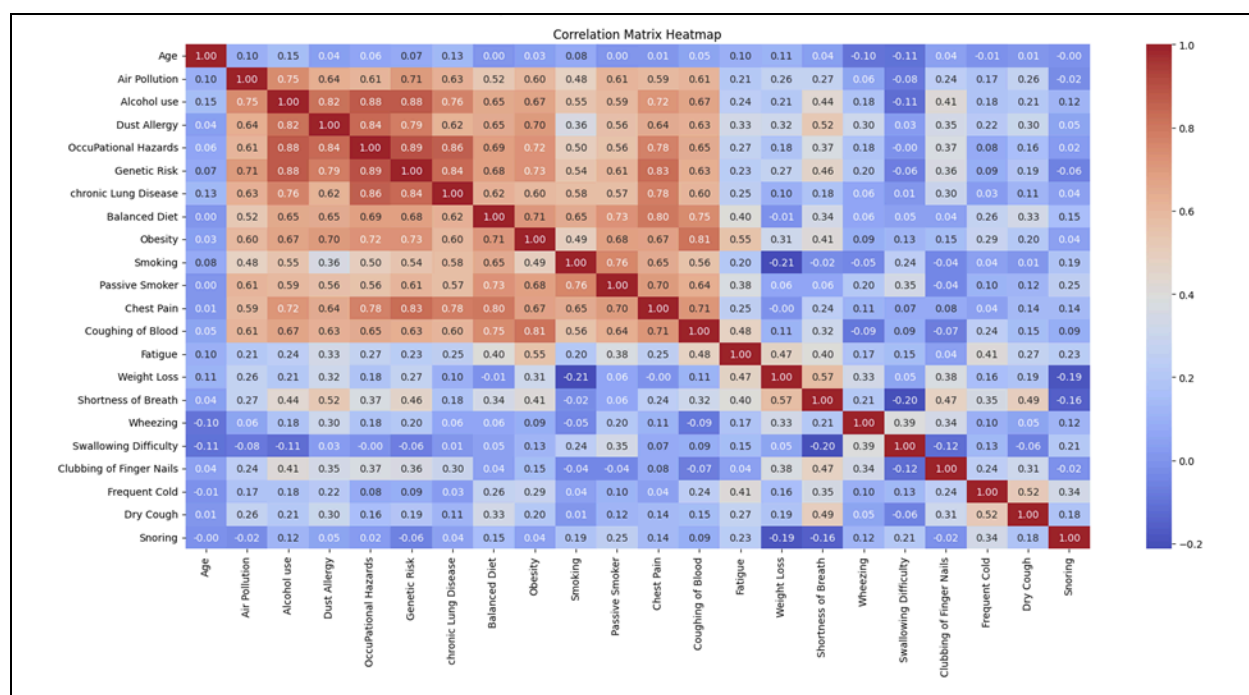


Fig. 10. This heat map displays the correlation matrix heatmap, with stronger associations in red and weaker associations in blue

General Pattern:

- From a quick glance of heatmap coloring, there is a clear group of 12 factors that form a red block of increased inter-factor correlations. These include air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, and coughing of blood. These align with the top factors from our factor versus severity level bar graph above.

Specific Findings:

- Air pollution is closely related to chronic lung disease, obesity, and smoking, indicating a potential role in their development.
- Smoking also shows strong links to coughing up blood, weight loss, and shortness of breath, highlighting its harmful health effects.
- Obesity has a strong correlation with a balanced diet, indicating a possible risk factor for certain conditions.
- Similar to our finding above, we strongly suspect “balanced diet” to be mislabelled. This is because a balanced diet is a good health behavior that should demonstrate protective effects against negative health outcomes. As such, we would expect a negative correlation between “balanced diet” and all of our other factors, especially obesity. However, we found a strong positive correlation of 0.71 between “balanced diet” and obesity, so we are quite confident that this data should be labeled “poor diet” instead.
- This heatmap provides valuable information for identifying health risks and developing strategies for prevention and treatment. It clearly shows how various factors are intertwined and can potentially impact our health.

Statistical Significance of Factor versus Factor Correlations

```
[ ] import pandas as pd
    from scipy.stats import pearsonr

    features_for_correlation = ['Age', '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']

    subset_df = df[features_for_correlation]

    correlations = subset_df.corr(method='pearson')

    for col1 in features_for_correlation[:-1]:
        for col2 in features_for_correlation[:-1]:
            if col1 != col2:
                correlation, p_value = pearsonr(subset_df[col1], subset_df[col2])
                print(f"Correlation between {col1} and {col2}: {correlation:.4f}, p-value: {p_value:.4f}")

    print("\nCorrelation Matrix:")
    print(correlations)
```

Fig. 11. Code for Spearman correlation between factors

We also calculated the statistical significance of the correlations between individual factors. Because our normality test found our data to not be normally distributed, we opted to measure correlation significance with the Spearman correlation coefficient. Due to the multitude of factors in our dataset, there are a plethora of correlation values calculated. Besides a few exceptions, the vast majority of p-values were less than 0.05, suggesting sufficient evidence to reject the null hypothesis and conclude that statistically significant correlations exist between nearly all factors. This finding suggests that our risk factors are not isolated. Rather than each factor being completely independent, they all seem to trend similarly. For example, an unhealthy person who smokes may also be likely to drink alcohol, have a poor diet, have dietary issues, and exhibit health problems.

5.4 DISTRIBUTION OF AGES (HISTOGRAM, PIE CHART, BAR GRAPH).



Fig. 12. This histogram displays the distribution of ages

The graph displaying the age distribution shows that the population is predominantly young, with the largest age group being individuals aged 30 - 39. As the age groups progress, the number of people in each age group gradually decreases, with the smallest group being individuals aged 70-79.

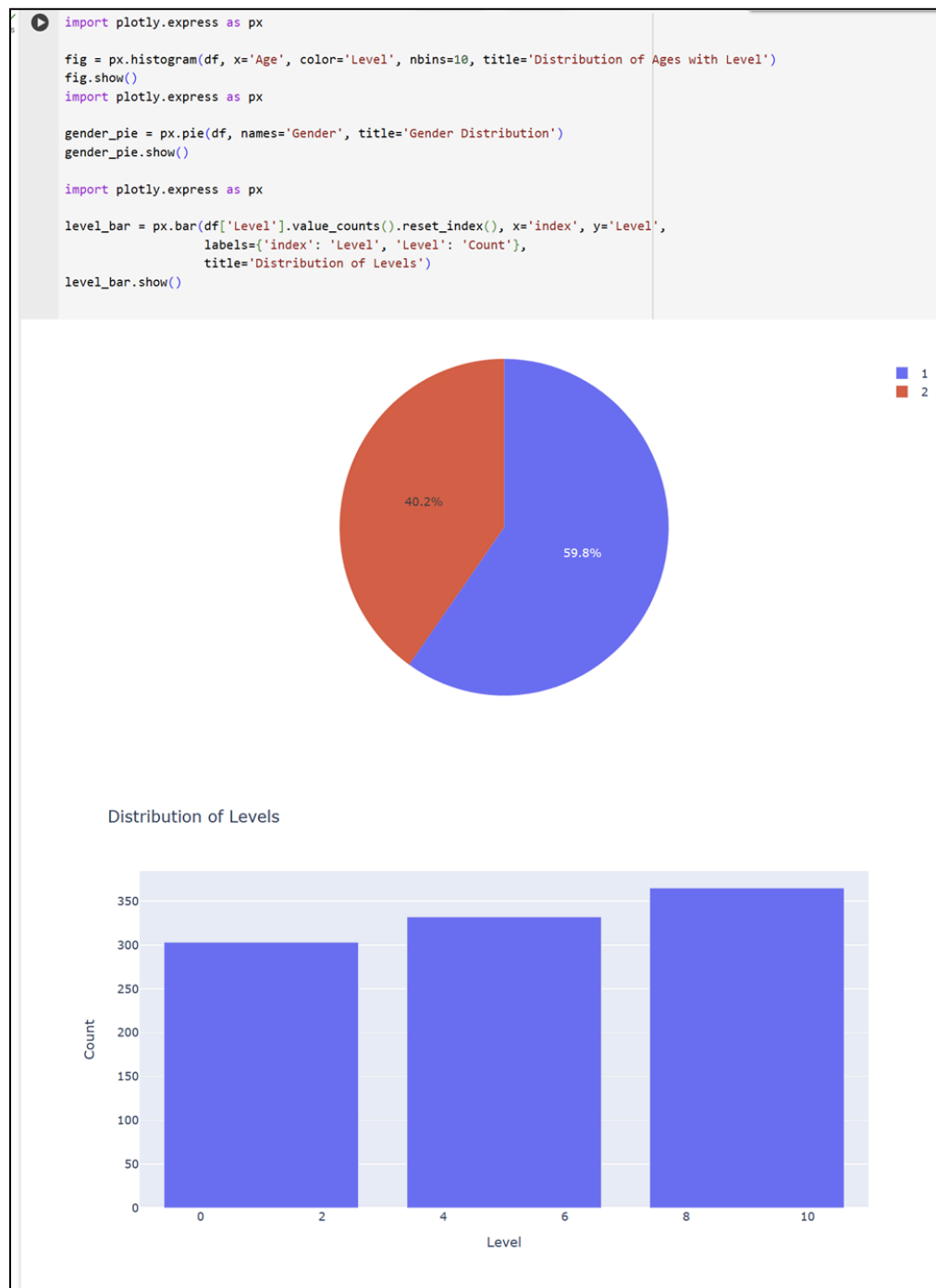


Fig. 13. shows the code, pie chart, and bar graph displaying level distribution

5.5 PATIENTS AND LEVEL DISTRIBUTION

PATIENTS COUNT AND LEVEL DISTRIBUTION

```
[25] gender_count = df['Gender'].value_counts()

      level_count = df['Level'].value_counts()

      print("Gender Distribution:")
      print(gender_count)

      print("\nLevel Distribution:")
      print(level_count)
```

```
Gender Distribution:
1    598
2    402
Name: Gender, dtype: int64

Level Distribution:
9    365
5    332
1    303
Name: Level, dtype: int64
```

Fig. 14. Code for the patients and level distribution

5.6 SCATTER PLOT

SCATTER PLOT

```
import plotly.express as px

scatter_plot = px.scatter(df, x='Age', y='Weight Loss', color='Level',
                          title='Scatter Plot of Age vs Weight Loss with Level')
scatter_plot.update_layout(title_x=0.5)
scatter_plot.show()
```



Scatter Plot of Age vs Weight Loss with Level

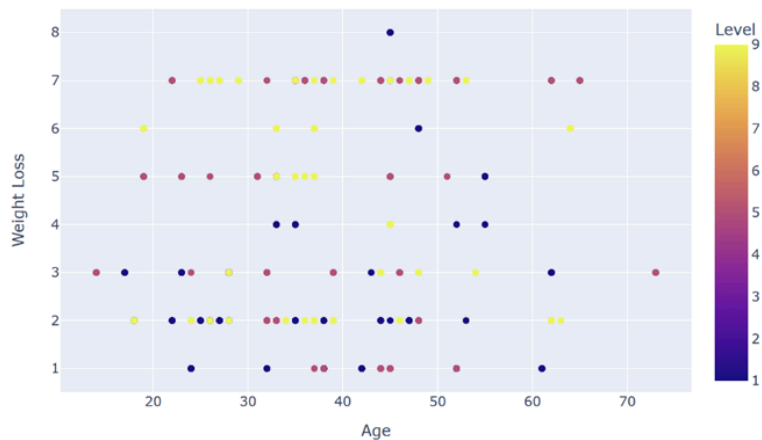


Fig. 15. Shows the code and scatter plot of age vs weight loss.

5.7 PIE CHART

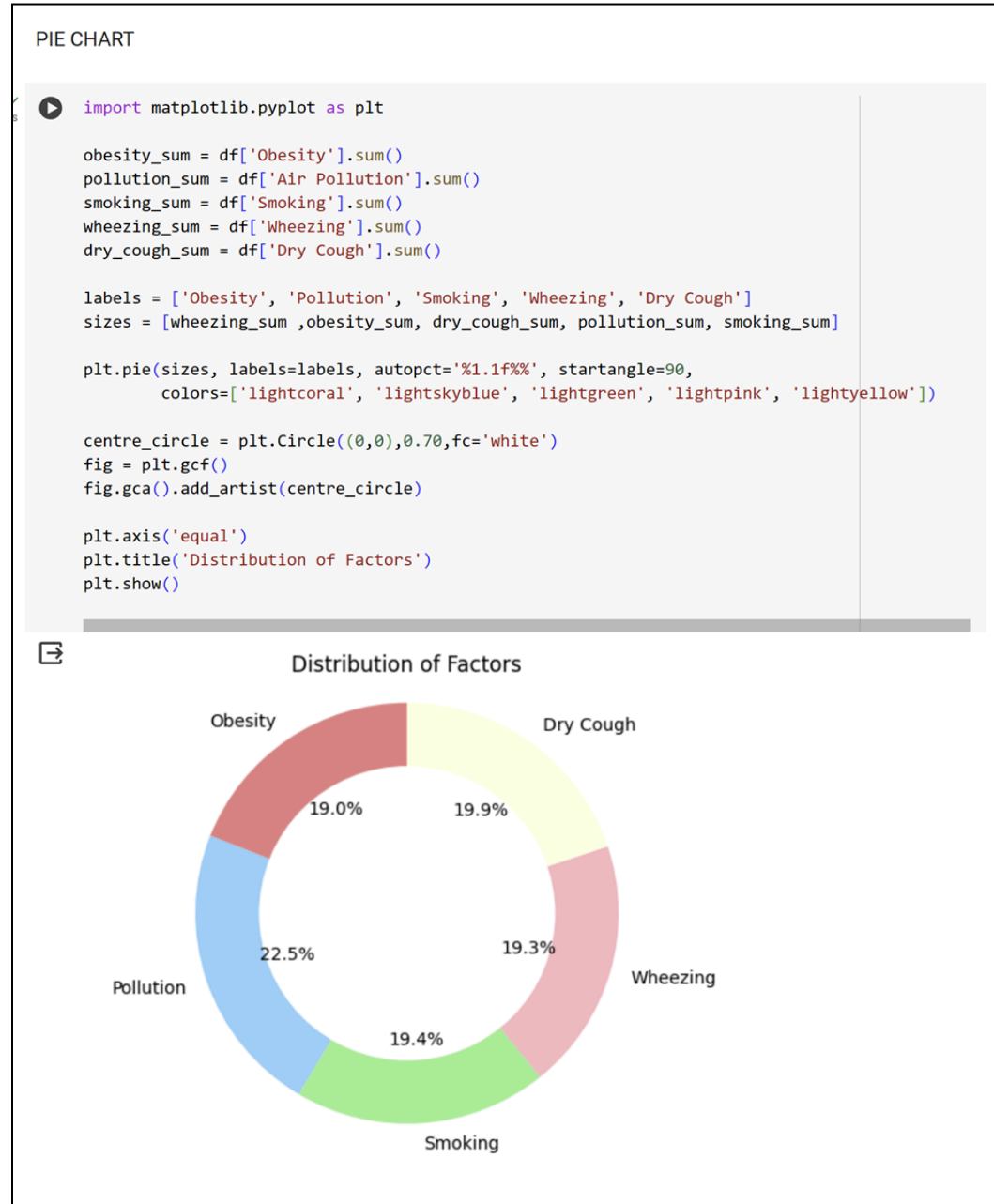


Fig. 16 The image displays a code alongside a pie chart, depicting the distribution of factors

The pie chart displays the percentage distribution of the following factors:

- Obesity (19.0%), Dry cough (19.9%), Wheezing (19.3%), Pollution (19.4%), Smoking (11.4%) . The pie chart illustrates that obesity, dry cough, wheezing, and pollution are the most prevalent factors, each accounting for approximately 19% of the total. Smoking is the least common factor, accounting for only 11.4% of the total.

5.8 PIE CHART FOR LUNG CANCER SEVERITY DISTRIBUTION

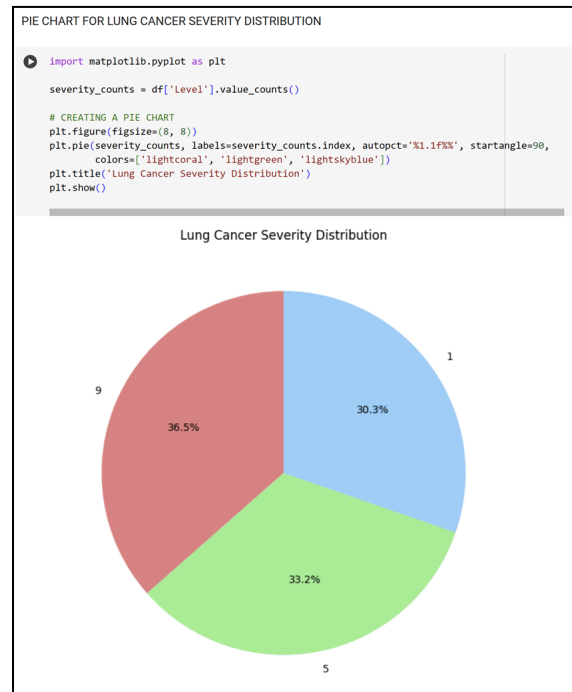


Fig. 17. Displays both the code and a pie chart that represents the distribution of lung cancer severity

5.8.1 PIE CHART MOST IMPACTFUL RISK FACTORS

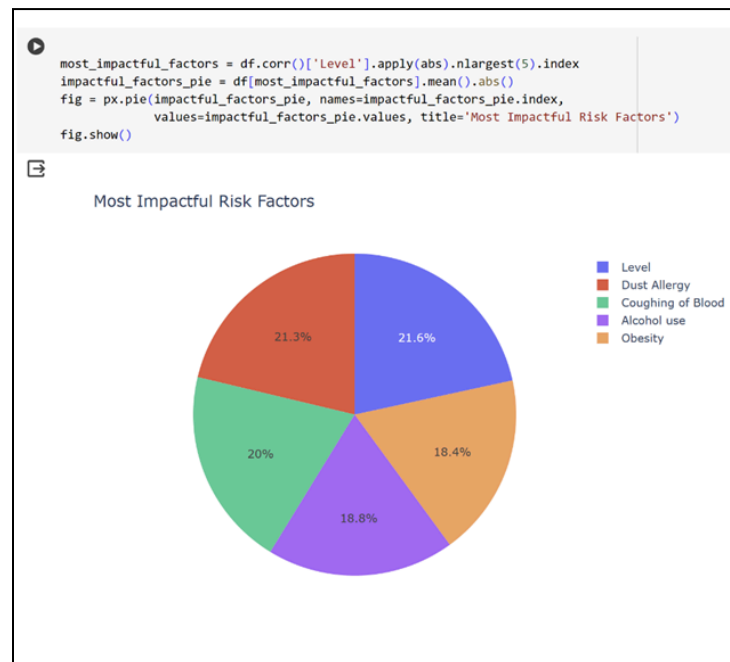


Fig. 18. Displays both the code and a pie chart that represents the most impactful Risk factors

5.9 BAR CHART FOR LUNG CANCER SEVERITY DISTRIBUTION

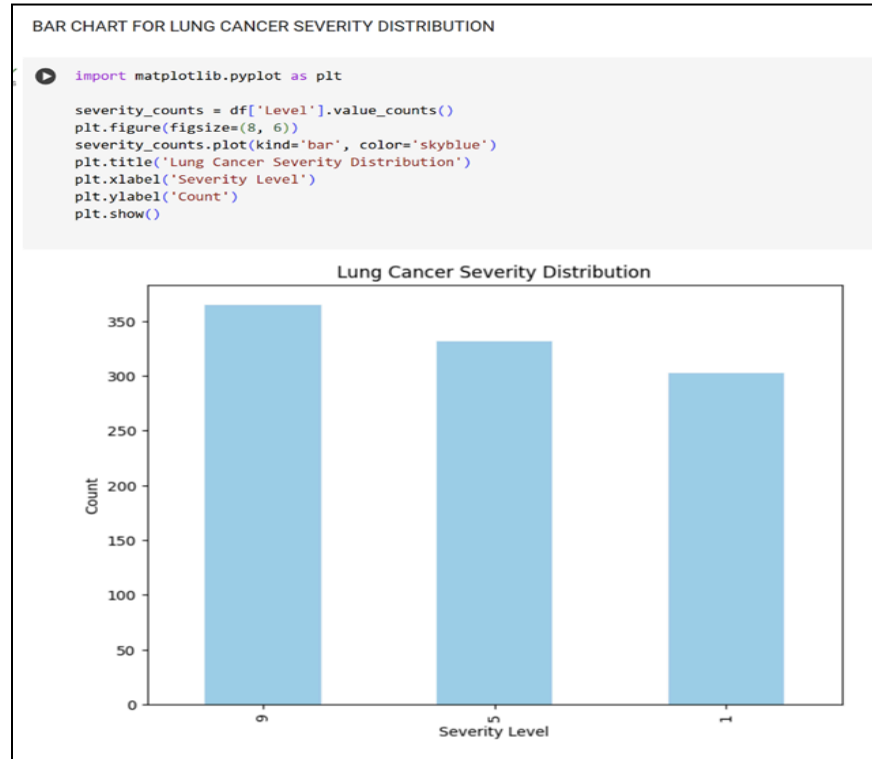


Fig. 19. Displays both the code and a Bar graph that represents the Lung Cancer Severity Distribution

5.10 SCATTER PLOT - (SUM OF FEATURES VS LUNG CANCER SEVERITY LEVEL)

SCATTER PLOT

```
import matplotlib.pyplot as plt

features_to_sum = ['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']

Sum_of_Features = df[features_to_sum].sum(axis=1)

figure(figsize=(10, 6))
scatter(df['Sum_of_Features'], df['Level'], color='blue', alpha=0.5)
title('Sum of Features vs Lung Cancer Severity Level')
xlabel('Sum of Features')
ylabel('Lung Cancer Severity Level')
grid(True)
show()
```

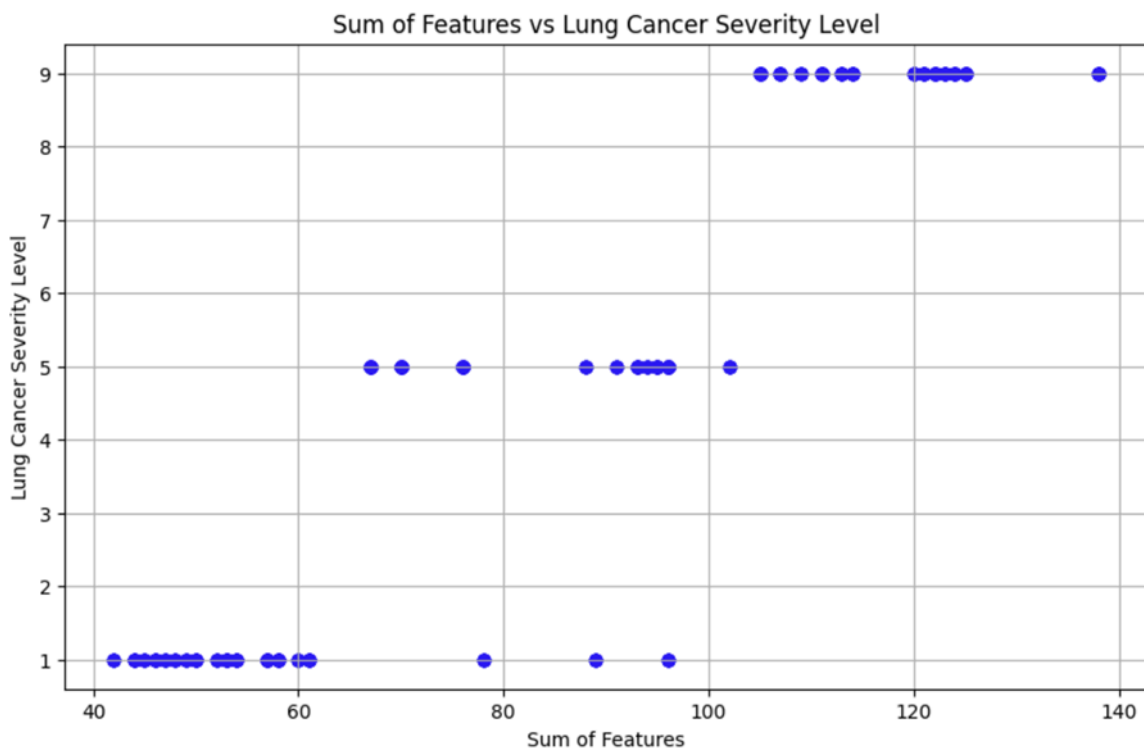


Fig. 20. Displays both the code and a Scatter plot that represents the Sum of features vs Lung cancer Severity Level

5.11 SCATTER PLOT (SCATTER PLOT OF AGE VS WEIGHT LOSS WITH COLOR CODED LEVELS)

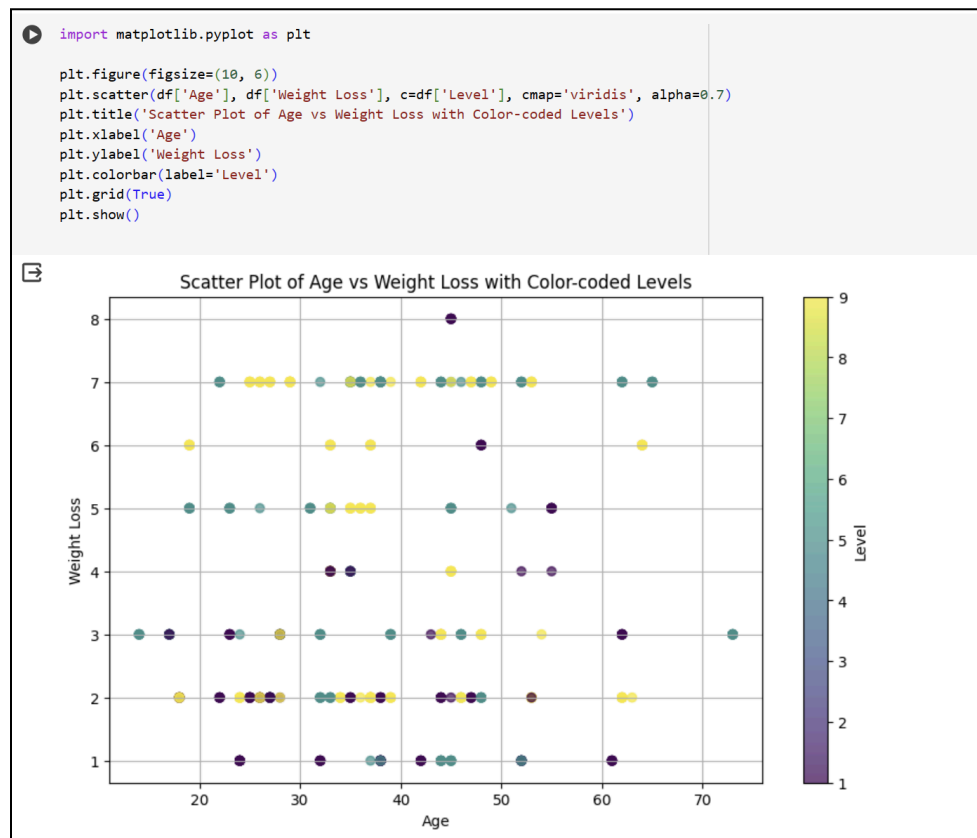


Fig. 21. Displays both the code and a Scatter plot that represents the Sum of Age vs Weight Loss with color coded levels

The graph displays the connection between age and weight loss. It indicates that with increasing age, individuals tend to lose less weight. However, there is considerable variation within each age group. While some older adults experience significant weight loss, some younger individuals may experience very little. The graph's color-coding represents the degree of weight loss, with red, orange, and yellow indicating higher levels of weight loss, and green, blue, and purple indicating lower levels of weight loss.

Several possible explanations would account for the overall trend of declining weight loss with age. For instance, slower metabolism and reduced muscle mass may make it more difficult for older people to lose weight. Another possibility is that older people are more likely to have medical conditions that can interfere with their weight loss efforts.

The variability in weight loss within each age group is likely due to several factors, such as diet, exercise, genetics, and other conditions. Some individuals may be more motivated to lose weight than others, and some may have more resources to support their weight loss endeavors. The graph illustrates a complicated relationship between age and weight loss. There is no single explanation for why some people lose weight and others do not. Nonetheless, the graph can assist us in developing hypotheses about the factors that influence weight loss in different age groups.

5.12 HISTOGRAM

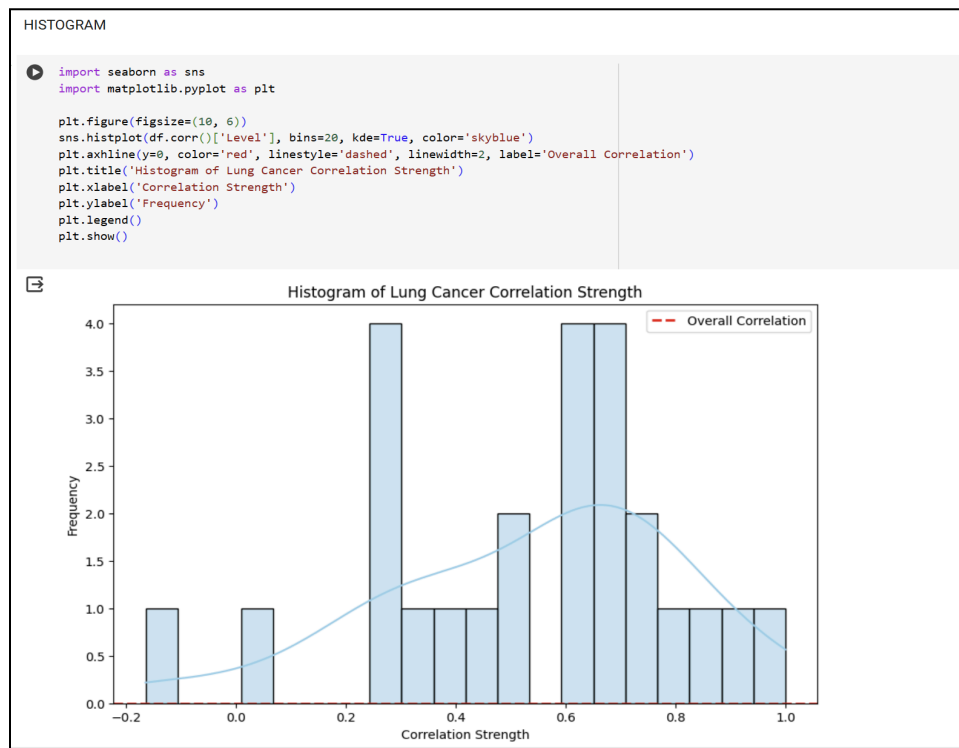


Fig. 22. Displays both the code and a Histogram that represents the Lung Cancer Correlation Strength

The histogram analysis indicates that most lung cancer risk factors are weakly to moderately correlated with each other. However, some risk factors, such as smoking, have stronger correlations with lung cancer risk. The overall correlation between all lung cancer risk factors is about 0.3, which means that having one risk factor increases the likelihood of having other risk factors. It is worth noting that correlation does not equal causation, but it can help identify potential risk factors for a disease.

The histogram shows that the most common correlation strength is between 0.2 and 0.4, suggesting that most lung cancer risk factors are weakly to moderately correlated. Smoking and lung cancer risk have a correlation strength of about 0.6, while other risk factors have weaker correlations. The dashed red line on the histogram represents the overall correlation between all lung cancer risk factors, which is positively correlated. This indicates that people with one risk factor are more likely to have other risk factors as well.

5.13 FOREST PLOT (FOREST PLOT OF ODDS RATIO WITH CONFIDENCE INTERVALS)

FOREST PLOT

```
[40] import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import statsmodels.api as sm

np.random.seed(42)
n_samples = 1000
features = ['Age', 'Air Pollution', 'Alcohol use', 'Obesity', 'Smoking', 'Genetic Risk', 'Chronic Lung Disease']
coefficients = [0.05, 0.2, -0.15, 0.1, -0.25, 0.1, -0.2]

data = pd.DataFrame(np.random.randn(n_samples, len(features)), columns=features)

data['Target'] = np.random.choice([0, 1], size=n_samples, p=[0.3, 0.7])

X = sm.add_constant(data[features])
logit_model = sm.Logit(data['Target'], X).fit()

odds_ratios = np.exp(logit_model.params)
conf_int = np.exp(logit_model.conf_int().iloc[:, 0]), np.exp(logit_model.conf_int().iloc[:, 1])

odds_ratios_df = pd.DataFrame({'Feature': X.columns[1:], 'Odds Ratio': odds_ratios.values[1:],
                              'Lower_CI': conf_int[0].values[1:], 'Upper_CI': conf_int[1].values[1:]})

odds_ratios_df = odds_ratios_df.sort_values(by='Odds Ratio', ascending=True)

plt.figure(figsize=(10, 6))
sns.barplot(x='Odds Ratio', y='Feature', data=odds_ratios_df, palette='viridis')
plt.errorbar(x=odds_ratios_df['Odds Ratio'], y=odds_ratios_df['Feature'],
             xerr=[odds_ratios_df['Odds Ratio'] - odds_ratios_df['Lower_CI'],
                   odds_ratios_df['Upper_CI'] - odds_ratios_df['Odds Ratio']],
             fmt='o', color='black', capsize=5, markersize=8)

plt.title('Forest Plot of Odds Ratios with Confidence Intervals')
plt.xlabel('Odds Ratio')
plt.ylabel('Feature')
plt.show()
```

Optimization terminated successfully.
Current function value: 0.593779
Iterations 5

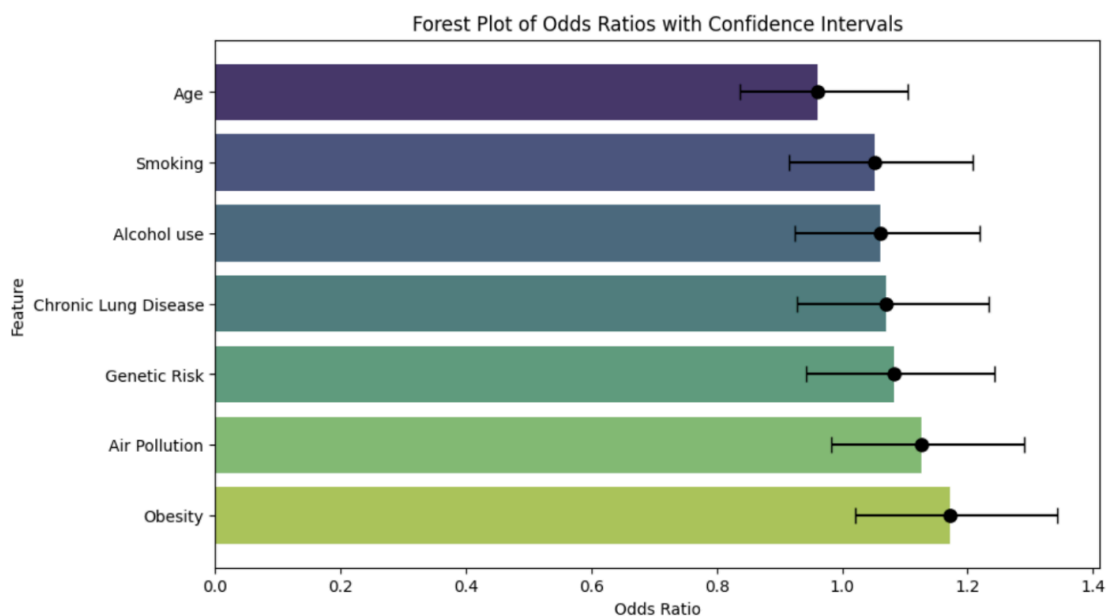


Fig. 22. Display both the code and a forest plot that represents the odds ratio with confidence intervals

Forest plots are visual representations that help us understand the relationship between various factors and the risk of developing chronic lung disease. From this forest plot, we can see that smoking poses the strongest risk factor for chronic lung disease. Its diamond stretches significantly to the right, indicating a substantial positive association. Air pollution, obesity, and genetic predisposition also appear to elevate disease risk, albeit to a lesser extent than smoking. Their diamonds extend slightly to the right, suggesting a measurable positive association. Other factors depicted in the plot show varying degrees of association with chronic lung disease, ranging from positive to negative. It's important to note that this forest plot presents adjusted estimates, meaning the influence of other factors has been taken into account. This allows for a clearer understanding of each factor's individual impact on disease risk.

Overall, the forest plot provides valuable insights into the various contributors to chronic lung disease. While smoking emerges as the most significant risk factor, other factors can also play a role. Understanding these associations can help individuals make informed choices and prioritize healthy lifestyle habits to mitigate disease risk.

6 Data Modeling

To utilize our dataset patterns to predict lung cancer risk, we constructed three distinct models: linear regression, decision tree classification, and k-nearest neighbors. This section briefly summarizes each model. The later discussion section further interprets each model and provides additional considerations for use in predicting risk.

6.1 Linear Regression

This model provides a continuous numerical prediction of risk on a 1-to-9 scale. This assumes linear relationships between factors and severity level.

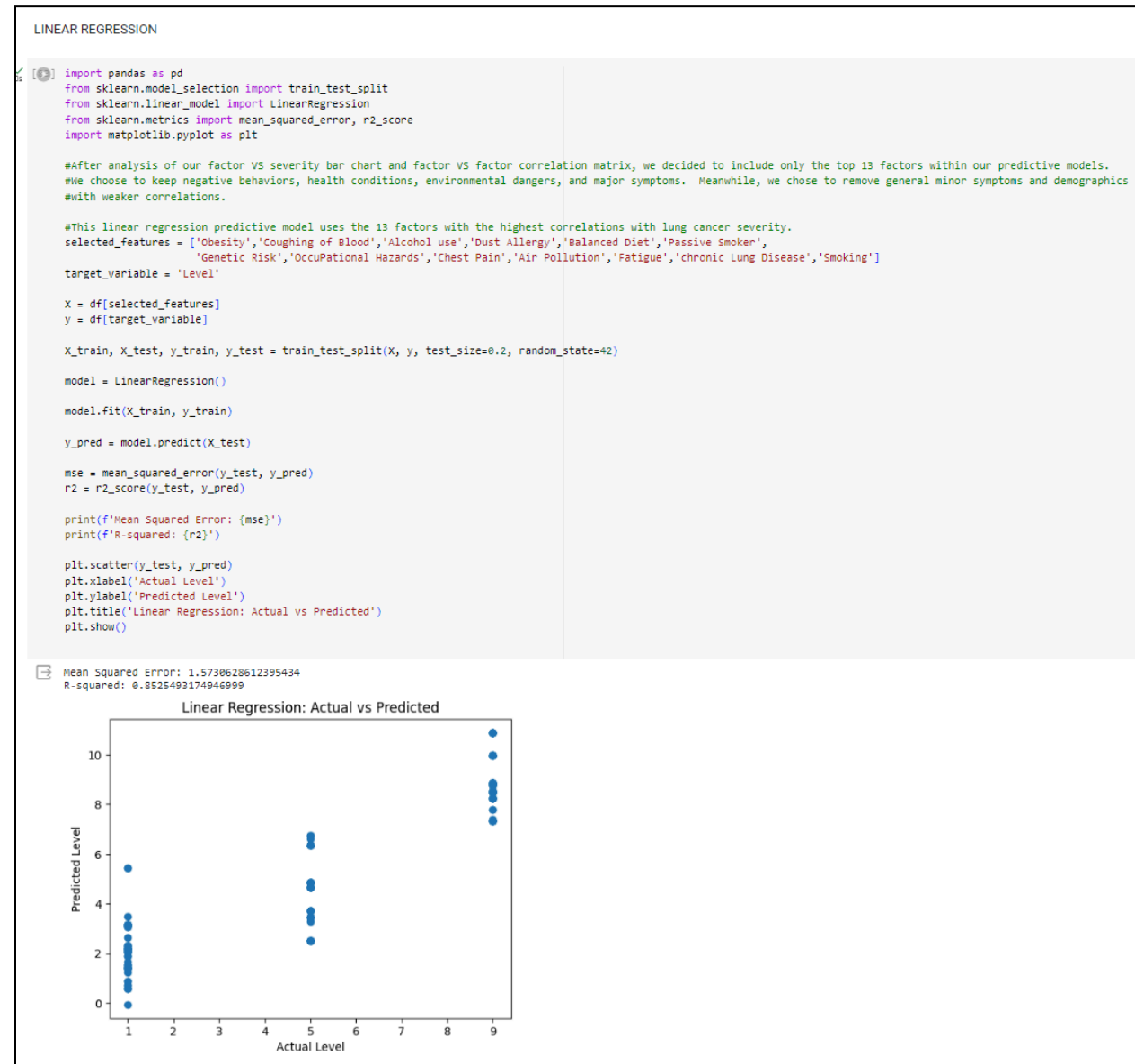


Fig. 23. Code for linear regression model with mean squared error and R-squared

6.2 Decision Tree Classifier

This categorical model assembles a hierarchy of factors. To predict severity level category, the model navigates through a series of decision nodes based on a new patient's factor data. This model constructs the hierarchy under the assumption that each factor has varying importance. The gini impurity measure can be used to explain the classification utility of a given node within the tree.

```
#Feature Importance using Decision Tree
features = ['Obesity','Coughing of Blood','Alcohol use','Dust Allergy','Balanced Diet','Passive Smoker',
            'Genetic Risk','OccuPational Hazards','Chest Pain','Air Pollution','Fatigue','chronic Lung Disease','Smoking']
target = 'Level'

df.dropna(subset=features + [target], inplace=True)

X = df[features]
y = df[target]

imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_imputed, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

dt_classifier = DecisionTreeClassifier(random_state=42)

dt_classifier.fit(X_train_scaled, y_train)

y_pred = dt_classifier.predict(X_test_scaled)

accuracy = accuracy_score(y_test, y_pred)
classification_report_result = classification_report(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'\nDecision Tree Classifier Score = {accuracy}')
print('\nClassification Report:\n', classification_report_result)
print('\nConfusion Matrix:\n', conf_matrix)

feature_importances = pd.DataFrame({'Feature': features, 'Importance': dt_classifier.feature_importances_})
print('\nFeature Importances:\n', feature_importances)
```

Decision Tree Classifier Score = 1.0

Classification Report:

	precision	recall	f1-score	support
1	1.00	1.00	1.00	55
5	1.00	1.00	1.00	63
9	1.00	1.00	1.00	82
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Confusion Matrix:

```
[[55  0  0]
 [ 0 63  0]
 [ 0  0 82]]
```

Feature Importances:

	Feature	Importance
0	Obesity	0.058371
1	Coughing of Blood	0.380759
2	Alcohol use	0.200131
3	Dust Allergy	0.044970
4	Balanced Diet	0.000000
5	Passive Smoker	0.021142
6	Genetic Risk	0.021370
7	OccuPational Hazards	0.071332
8	Chest Pain	0.023097
9	Air Pollution	0.000000
10	Fatigue	0.153480
11	chronic Lung Disease	0.025349
12	Smoking	0.000000

Fig. 24. Code for decision tree classifier model with accuracy measurements and feature importance

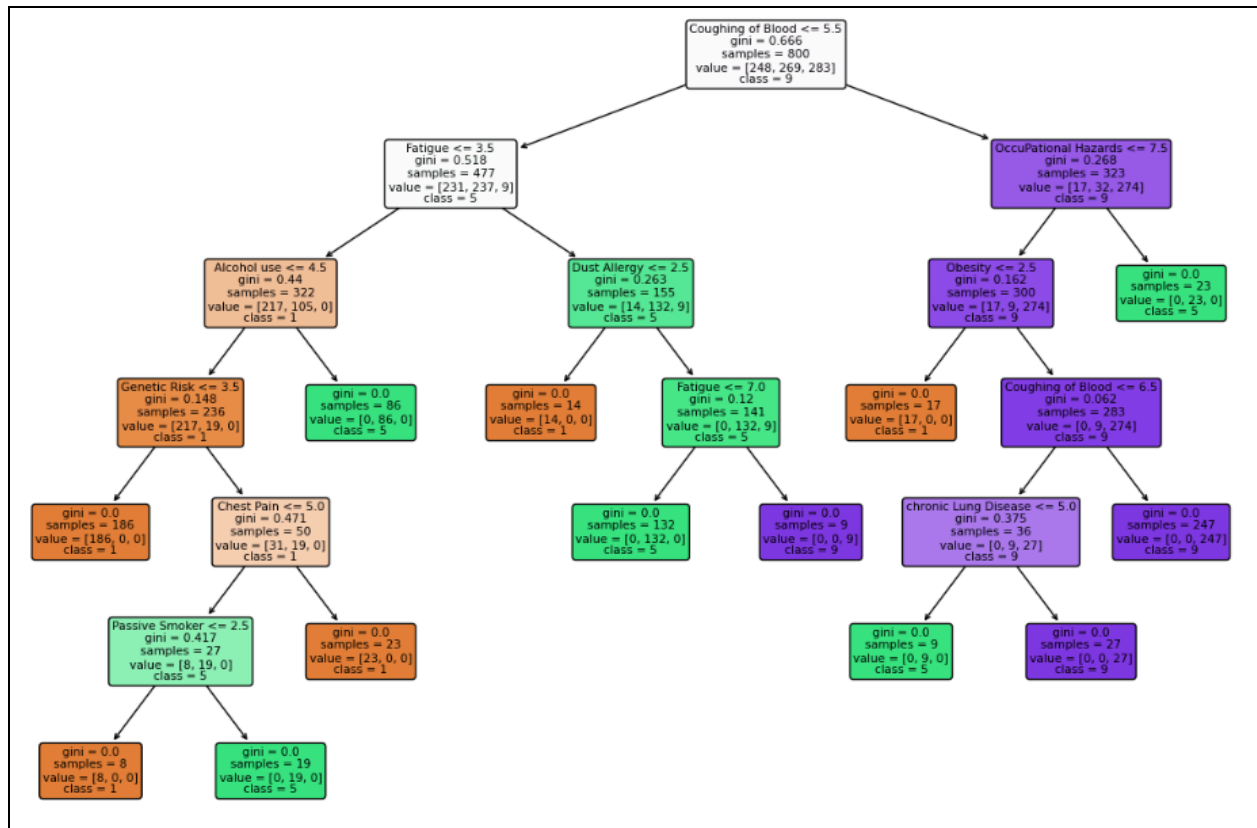


Fig. 25. Decision tree model visualization

6.3 K-nearest Neighbors

This categorical model predicts the severity level category of new patients based on the 7 dataset patients with the most similar data to the new patient. This model assumes that when factor data is similar, severity level is also similar.

K NEAREST NEIGHBOR

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

#K nearest neighbors
features = ['Obesity', 'Coughing of Blood', 'Alcohol use', 'Dust Allergy', 'Balanced Diet', 'Passive Smoker',
           'Genetic Risk', 'Occupational Hazards', 'Chest Pain', 'Air Pollution', 'Fatigue', 'Chronic Lung Disease', 'Smoking']
target = 'Level'

X = df[features]
y = df[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


k = 7
knn_classifier = KNeighborsClassifier(n_neighbors=k)

knn_classifier.fit(X_train_scaled, y_train)

y_pred = knn_classifier.predict(X_test_scaled)

accuracy = accuracy_score(y_test, y_pred)
classification_report_result = classification_report(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'\nK-Nearest Neighbors Classifier Score (k={k}): {accuracy}')
print('\nClassification Report:\n', classification_report_result)
print('\nConfusion Matrix:\n', conf_matrix)
```

 K-Nearest Neighbors Classifier Score (k=7): 1.0

Classification Report:

	precision	recall	f1-score	support
1	1.00	1.00	1.00	55
5	1.00	1.00	1.00	63
9	1.00	1.00	1.00	82
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Confusion Matrix:

```
[[55  0  0]
 [ 0 63  0]
 [ 0  0 82]]
```

Fig. 26. Code for K-nearest neighbors model with accuracy measurements

6.4 Factor Inclusion

During feature representation, we confirmed in scientific literature that all factors were potentially relevant to lung cancer. However, our data analysis and visualization results suggested that not all factors contribute to risk equally. As such, we experimented with several combinations to assess the optimal factors to include in our models. During testing, we initially included all 23 factors within our models but were surprised and by the concerning perfect accuracy:

```
Decision Tree Classifier Accuracy: 1.0  
Linear Regression R-squared Score: 0.9291587235558918  
K-Nearest Neighbors Classifier Accuracy: 1.0
```

Fig. 27. Evaluation of models that include all 23 factors

These accuracy scores demonstrated potential risk of overfitting, especially for the categorical models. So, we also tested the models with only including 5 completely random factors and still found near-perfect accuracy with our categorical models:

```
Decision Tree Classifier Accuracy: 0.97  
Linear Regression R-squared Score: 0.5957691869482428  
K-Nearest Neighbors Classifier Accuracy: 0.97
```

Fig. 28. Evaluation of models that include 5 random factors

To guide our factor inclusion, we noted a group of factors within our correlation matrix heatmap that shows strong correlations with each other and clearly forms a red block. Based on our bar chart, these factors also showed the greatest factor versus severity correlation strengths. For our models, we decided to include the top 13 factors from our factor versus severity correlation data, with smoking as the cutoff factor. This is because smoking is widely emphasized as the most crucial risk factor within scientific literature. On our correlation strength list, most factors after smoking were generally mild health symptoms that did not clearly pertain to lung cancer. By omitting the least impactful factors, we reduce the risk that our models overfit our dataset and increase generalizability for risk prediction of new patients.

```
Decision Tree Classifier Accuracy: 1.0  
Linear Regression R-squared Score: 0.8525493174946999  
K-Nearest Neighbors Classifier Accuracy: 1.0
```

Fig. 29. Evaluation of models that include the top 13 factors based on correlation strength with severity level

7 Results

Our data analysis, visualization, and modeling procedures provided us valuable insights to address our two study aims.

7.1 Aim 1

Our first aim was to determine whether any risk factors are correlated with lung cancer severity. Findings relevant to this aim include:

- Statistically significant correlations were found between each risk factor and severity level.
- From strongest to weakest, the order of factor correlation strength with severity level in our dataset is: obesity, coughing of blood, alcohol use, dust allergy, balanced diet, passive smoker, genetic risk, occupational hazards, chest pain, air pollution, fatigue, chronic lung disease, smoking, shortness of breath, frequent cold, dry cough, weight loss, snoring, clubbing of fingernails, swallowing difficulty, wheezing, gender, and age.

7.2 Aim 2

Our second aim was to identify interactions between risk factors. Findings relevant to this aim include:

- Statistically significant correlations were found between nearly all risk factors.
- The correlation matrix heatmap clearly indicated a set of factors with increased associations with each other. In no particular order, these included: air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, and fatigue. This inter-factor pattern aligned with our top 13 factors based on correlation strength with severity level.

7.3 Aim 3

Our third aim was to construct models that predict lung cancer risk. Findings relevant to this aim include:

- The k-nearest neighbors and decision tree models provided categorical risk predictions
 - high accuracy (add specific measurements)
- The linear regression model provided continuous numerical risk predictions.
 - moderate-to-high accuracy
- Model accuracy increases as the number of included factors increases.

8 Discussion

For our project, we were able to meet our goals to assess factor versus severity correlations, factor versus factor patterns, and create predictive models. In analysis, we confirmed statistically significant correlations and rejected our null hypothesis. However, in interpreting our results, we note several key considerations.

8.1 Comparison to Scientific Literature

Firstly, we have concerns regarding the pattern discrepancies between our dataset and existing scientific literature. The American Cancer Society (2023) notes that 80% of lung cancer deaths are associated with smoking. Secondhand smoke, pollution, and carcinogen exposure are other key factors. With this knowledge, we expected smoking, passive smoking, occupational hazards, and air pollution show the greatest correlation strengths with severity level. On one hand, we did find positive correlations between all our factors and severity.

On the other hand, we found several strange patterns. Firstly, smoking ranked 13th, which was among the bottom 50% of risk factor influence. Furthermore, passive smoking ranked 6th, ahead of smoking itself. This doesn't make logical sense, because someone who directly smokes would be expected to face higher lung cancer risk than someone who breathes in secondhand smoke. Thirdly, unexpected factors such as obesity, coughing of blood, alcohol use, dust allergy, and balanced diet were ranked unusually high. While such factors were certainly expected to have positive correlations with severity level, they weren't expected to rank higher than the primary risk factors from scientific literature. As a result of these patterns, our dataset seems to have questionable validity.

8.2 Model Evaluation

Though we have measured accuracy of our models, there are additional considerations in evaluating the utility of models.

Linear regression. Relative to the categorical models, linear regression did not feature perfect accuracy. However, a major positive is that this model provides granularity, where risk is evaluated on a continuous numerical scale from 1 to 9, contrasting the limited 3-point prediction of our categorical models. If our models were used in practice, a risk assessment of "medium" may not provide any meaningful understanding of risk, especially when the criteria for assigning categories is not known. In comparison, a risk assessment of "3.83 on a scale of 1 to 9" could provide clinicians and patients with a more precise idea of risk.

Decision tree classifier. Accuracy measures found this model to be extremely accurate, even when only 5 factors were included. However, considering data type and complexity, a decision tree may not have been the most applicable model due to unclear factor priority and splitting criteria. As a comparison, identifying the species of an unknown organism is a situation where a decision tree would be optimal, because there are clear factor priorities from broadest to narrowest, such as single-cell or multicellular, backbone or no backbone, and number of legs.

However, our data contains a combination of factors that do not necessarily follow a hierarchy. Regardless of applicability, the decision tree classifier algorithm will still create a hierarchy of our factors. As a result, a decision tree can lead to overfitting. Furthermore, predicting severity level among three categories is easier than predicting a continuous numerical value as our linear regression model provides.

K-nearest neighbors. Similar to our other categorical model, KNN was also found to be extremely accurate but faces similar challenges as the decision tree classifier model. Predictions are entirely based on matching a new patient with the closest 7 patients from our dataset. For model predictions to be useful, it must be assumed that the training dataset is accurate. While 1000 is a relatively large sample size, any deviations between our sample and the true population would result in skewed predictions that overfit our sample. Furthermore, this model is also only capable of providing predictions that fall into three categories.

8.3 Considerations For Use

In addition to the model-specific interpretations above, we note several considerations regarding practical use of such models in predicting lung cancer risk.

Accuracy. In our code, we evaluated the accuracy of our models using various metrics such as accuracy, precision, recall, and F1-score. Additionally, for the linear regression model, we also calculated the mean squared error and r-squared values. Both categorical models exhibited exceptional accuracy. When including only 5 factors in the model, accuracy already reached 0.97. Including 13 factors allowed the categorical models to reach 100% accuracy. The continuous linear regression model had an accuracy ranging from 0.60 to 0.93 depending on the number of factors included in the model. Though categorical model accuracy scores are higher, there are two major caveats. Firstly, our categorical models can correctly guess much more easily when only 3 categorical possibilities exist. In contrast, our continuous regression model has to choose between all numerical possibilities from 1 to 9. Secondly, such high accuracy scores can indicate overfitting, where models that align too closely with training data do not perform as effectively when presented with new data.

Validity. The computer science phrase “garbage in, garbage out” describes the phenomenon that even if proper steps are taken in data analysis, visualization, and modeling, the utility of results is jeopardized if the original input is invalid. Based on the discussed pattern discrepancies between our dataset and scientific literature, the phrase may apply to our models. Even though we constructed several models to predict severity and verified their accuracy, models may not be practically useful in predicting lung cancer if our dataset does not contain accurate measurements of a true lung cancer population.

Generalizability. To evaluate risk factors for preventative medicine, a more ideal dataset for analysis would include disease presence (rather than disease severity level) as the dependent variable. This way, patterns could be analyzed to compare groups based on presence or absence of lung cancer. Instead, all 1000 of our patients had lung cancer. For predicting lung cancer risk in the general population where most people do not have lung cancer, generalizability of our models may be limited.

Practical Predictive Utility. While all factors have correlations have severity, they do not all have the same usefulness in predictive models. Firstly, categorical models are only constrained to three categories, which limits the granularity of predictions. Also, the dataset author does not specify the exact criteria used to distinguish between the severity levels. Additionally, for the goal of facilitating preventative medicine, symptom factors may have limited utility in a predictive model. For example, our decision tree model indicated coughing of blood as a key factor, but such a severe symptom may already indicate development of advanced disease. As such, predictive models may be more useful if based on adjustable behaviors or environmental conditions.

9 Limitations

- Dependence on data from medical records. The recorded patient risk factors and coexisting medical conditions might contain errors or discrepancies. Furthermore, significant confounding factors might not exist.

- A limited ability to generalize. Only individuals with lung cancer were included in the research. For screening purposes, the results and predictive model might not be applicable to the broader population.
- No cohort for external validation. The same dataset will be used for both training and testing the model. No impartial dataset exists to verify performance. One risk is overfitting.
- Linear relationships are assumed. The application of linear regression modeling makes the assumption that risk factors and the severity of lung cancer have linear connections; nevertheless, certain interactions may not be linear.
- Limitations on self-reported data. Subject self-report is the basis for several variables, such as smoking and drinking history, which may be skewed or unreliable.
- The size of the sample. Even with a large sample size of 1,000 patients, tiny subgroups may still impair the prediction accuracy and reliability of multivariate regression modeling.
- Bias in classification. Clinical interpretation during the data collecting step is necessary for the three-level classification of cancer severity: low, medium, and high. Misclassification within categories is a possibility.
- Dataset is from 2017, which is not current and may not reflect lung cancer risk patterns today.
- Entire group comes from purely health backgrounds and were beginners in coding and data science.

10 Appendix

Full Python ipynb file will be submitted separately. A few updates have been made to the code since the initial presentation code submission on 04-Dec.

11 References

- Ahmad, A. S., & Mayya, A. M. (2020). A new tool to predict lung cancer based on risk factors. *Heliyon*, 6(2), e03402. <https://doi.org/10.1016/j.heliyon.2020.e03402>
- American Cancer Society. (2023). About lung cancer. <https://www.cancer.org/cancer/types/lung-cancer/about.html>
- Cleveland Clinic. (2023). Nail clubbing. <https://my.clevelandclinic.org/health/symptoms/24474-nail-clubbing>
- Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C., & Boffetta, P. (2016). Risk factors for lung cancer worldwide. *The European respiratory journal*, 48(3), 889–902. <https://doi.org/10.1183/13993003.00359-2016>
- May, L., Shows, K., Nana-Sinkam, P., Li, H., & Landry, J. W. (2023). Sex Differences in lung cancer. *Cancers*, 15(12), 3111. <https://doi.org/10.3390/cancers15123111>
- Prithivraj. (2017). Lung cancer data. <https://data.world/cancerdatahp/lung-cancer-data>