Project Report

Predictive Analysis of Factors Associated with Lung Cancer Incidence

Presented by : GROUP-6
Anil Rai Kandimamalla, Babburi soumith sai,
Divya Sri Mukku, Hima Bindhu Sivva, Rohitha Ganta, Vahini Matlakunta

I. INTRODUCTION

In this study we delve into the realm of lung cancer prediction going beyond the well-known risk factor of smoking. Our research focuses on analyzing a dataset that considers not only lifestyle choices but also environmental factors with a specific emphasis on air pollution's role. The primary goal is to enhance the accuracy of lung cancer risk assessments and customize treatment plans effectively based on each patient's profile, especially considering the growing global prevalence of lung cancer. By exploring this data our aim is to shed light on explored causes of lung cancer and potentially uncover new risk factors. This endeavor is crucial for advancing medicine in oncology providing hope for improved prognostic and treatment approaches. Our study represents an effort in comprehending the origins of lung cancer that is expected to significantly enhance patient outcomes and offer new avenues for effective treatment.

II. DESCRIPTION OF VARIABLES

- Age: The age of the patient. (Numeric)
- Gender: The gender of the patient. (Categorical)
- Air Pollution: The level of air pollution exposure of the patient. (Categorical)
- Alcohol use: The level of alcohol use of the patient. (Categorical)
- Dust Allergy: The level of dust allergy of the patient. (Categorical)
- Occupational Hazards: The level of occupational hazards of the patient. (Categorical)
- Genetic Risk: The level of genetic risk of the patient. (Categorical)
- Chronic Lung Disease: The level of chronic lung disease of the patient. (Categorical)
- Balanced Diet: The level of a balanced diet of the patient. (Categorical)
- Obesity: The level of obesity of the patient. (Categorical)
- Smoking: The level of smoking of the patient. (Categorical)
- Passive Smoker: The level of passive smoker of the patient. (Categorical)
- Chest Pain: The level of chest pain of the patient. (Categorical)
- Coughing of Blood: The level of coughing of blood of the patient. (Categorical)
- Fatigue: The level of fatigue of the patient. (Categorical)
- Weight Loss: The level of weight loss of the patient. (Categorical)
- Shortness of Breath: The level of shortness of breath of the patient. (Categorical)
- Wheezing: The level of wheezing of the patient. (Categorical)
- Swallowing Difficulty: The level of swallowing difficulty of the patient. (Categorical)
- Clubbing of Fingernails: The level of clubbing of the fingernails of the patient. (Categorical)

II. DATA EXPLORATION

We have utilized Kaggle to download our dataset:
https://www.kaggle.com/datasets/thedevastator/cancer patients
The dataset contains 1000 rows and 26 columns.

To understand the structure of our dataset we can utilize the str() function. This handy function offers an overview of the datasets organization showcasing the types initial entries, in each column and overall dimensions of the dataset.

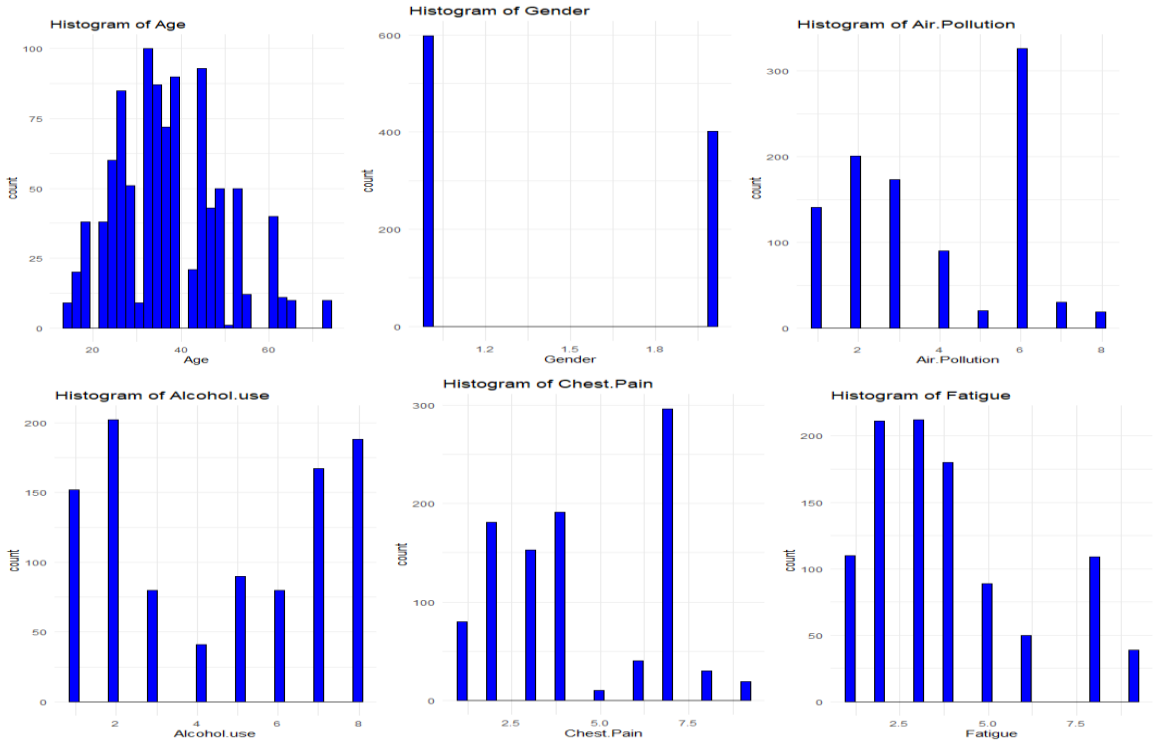
```
> str(df)
'data.frame':   1000 obs. of  26 variables:
 $ index                   : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Patient.Id              : chr  "P1" "P10" "P100" "P1000" ...
 $ Age                     : int  33 17 35 37 46 35 52 28 35 46 ...
 $ Gender                  : int  1 1 1 1 1 1 2 2 2 1 ...
 $ Air.Pollution           : int  2 3 4 7 6 4 2 3 4 2 ...
 $ Alcohol.use             : int  4 1 5 7 8 5 4 1 5 3 ...
 $ Dust.Allergy            : int  5 5 6 7 7 6 5 4 6 4 ...
 $ OccuPational.Hazards    : int  4 3 5 7 7 5 4 3 5 2 ...
 $ Genetic.Risk            : int  3 4 5 6 7 5 3 2 6 4 ...
 $ chronic.Lung.Disease    : int  2 2 4 7 6 4 2 3 5 3 ...
 $ Balanced.Diet           : int  2 2 6 7 7 6 2 4 5 3 ...
 $ Obesity                 : int  4 2 7 7 7 7 4 3 5 3 ...
 $ Smoking                 : int  3 2 2 7 8 2 3 1 6 2 ...
 $ Passive.Smoker          : int  2 4 3 7 7 3 2 4 6 3 ...
 $ Chest.Pain              : int  2 2 4 7 7 4 2 3 6 4 ...
 $ Coughing.of.Blood       : int  4 3 8 8 9 8 4 1 5 4 ...
 $ Fatigue                 : int  3 1 8 4 3 8 3 3 1 1 ...
 $ Weight.Loss             : int  4 3 7 2 2 7 4 2 4 2 ...
 $ Shortness.of.Breath     : int  2 7 9 3 4 9 2 2 3 4 ...
 $ Wheezing                : int  2 8 2 1 1 2 2 4 2 6 ...
 $ Swallowing.Difficulty   : int  3 6 1 4 4 1 3 2 4 5 ...
 $ Clubbing.of.Finger.Nails: int  1 2 4 5 2 4 1 2 6 4 ...
 $ Frequent.Cold           : int  2 1 6 6 4 6 2 3 2 2 ...
 $ Dry.Cough               : int  3 7 7 7 2 7 3 4 4 1 ...
 $ Snoring                 : int  4 2 2 5 3 2 4 3 1 5 ...
 $ Level                   : chr  "Low" "Medium" "High" "High" ...
```

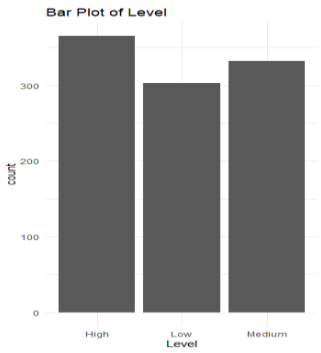
Overview of the Dataset:

We will be using head() to look into the first six rows of the data

```
> head(df)
  index Patient.Id Age Gender Air.Pollution Alcohol.use Dust.Allergy OccuPational.Hazards Genetic.Risk chronic.Lung.Disease Balanced.Diet Obesity Smoking
1     0         P1  33      1             2           4            5                    4            3                    2             2       4       3
2     1        P10  17      1             3           1            5                    3            4                    2             2       2       2
3     2       P100  35      1             4           5            6                    5            5                    4             6       7       2
4     3      P1000  37      1             7           7            7                    7            6                    7             7       7       7
5     4       P101  46      1             6           8            7                    7            7                    6             7       7       8
6     5       P102  35      1             4           5            6                    5            5                    4             6       7       2
  Passive.Smoker Chest.Pain Coughing.of.Blood Fatigue Weight.Loss Shortness.of.Breath Wheezing Swallowing.Difficulty Clubbing.of.Finger.Nails Frequent.Cold Dry.Cough
1              2          2                 4       3           4                   2        2                      3                      1              2         3
2              4          2                 3       1           3                   7        8                      6                      2              1         7
3              3          4                 8       8           7                   9        2                      1                      4              6         7
4              7          7                 8       4           2                   3        1                      4                      5              6         7
5              7          7                 9       3           2                   4        1                      4                      2              4         2
6              3          4                 8       8           7                   9        2                      1                      4              6         7
  Snoring  Level
1       4    Low
2       2 Medium
3       2   High
4       5   High
5       3   High
6       2   High
```

Visualizations:





III. METHODS & RESULTS:

Research Question 1: Are there differences, in the age distribution, among female patients?

To address this question we will utilize a method called the t test. The t-test allows us to compare the age of female patients.

The calculated test statistic is 6.8571. The t value gauges the magnitude of the difference, in relation to the variability observed in your sample data. A higher t-value suggests a distinction between the groups.

Based on the provided sample estimates we can observe that the average value of x (representing males) is 39.16221 while the average value of y (representing females) is 34.21642.

Based on the 95% confidence interval it can be determined that the actual variance, in age, between males and females falls within the range of 3.5308383 to 6.631196.

```
male_ages <- df$Age[df$Gender == 1]
female_ages <- df$Age[df$Gender != 1]

t_test_result <- t.test(male_ages, female_ages)

print(t_test_result)
```

```
        Welch Two Sample t-test

data:  male_ages and female_ages
t = 6.8571, df = 980.42, p-value = 1.242e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.530383 6.361196
sample estimates:
mean of x mean of y
 39.16221  34.21642
```

2. Is there a difference in air pollution exposure based on smoking levels?

To tackle this question we will utilize an ANOVA analysis. We chose the ANOVA test as it enables us to effectively explore the differences, in variability within and, between groups based on smoking habits thereby facilitating comparisons of the means.

Null Hypothesis: There is no difference in air pollution exposure based on smoking levels.

Alternative Hypothesis:There is a difference in air pollution exposure based on smoking levels.

```
df$Smoking <- as.factor(df$Smoking)

result <- aov(df$Air.Pollution ~ df$Smoking, data = df)

# Get the summary of the ANOVA test
summary(result)
```

```
> summary(result)
             Df Sum Sq Mean Sq F value Pr(>F)
df$Smoking    7   2344   334.9   187.3 <2e-16 ***
Residuals   992   1774     1.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The smoking variable has 7 degrees of freedom (df) indicating that there are 8 levels of smoking, as the number of groups minus 1. The F value of 187.3 reveals the extent to which air pollution levels differ among smoking levels compared to how much they vary within each smoking level. Given its value it suggests a substantial impact is present.

The calculated p-value of 2.2e-16 being, then 0.05 indicates a level of statistical significance. This implies that there is a variation in the levels of air pollution individuals are exposed to depending on their smoking habits.

The data strongly suggests that there is an association between smoking habits and the amount of air pollution and encounters.

3. Can we use age, gender and air pollution as predictors to determine the severity of coughing up blood?

To tackle this question we will apply Logistic Regression as our chosen method. We opted for regression because the data involves rankings. Furthermore logistic regression is suitable since our dependent variable is categorical, in nature.

Null Hypothesis: There is no relationship between age, gender, air pollution and the severity of coughing blood.

Alternative Hypothesis: There is a relationship between age, gender, air pollution and the severity of coughing blood.

The positive coefficient (0.002929) for Age suggests a link to coughing blood. However with a t value of 0.6013 this relationship lacks significance at a significance level of 0.05. Conversely the negative coefficient (0.1154) for Gender 2 (female) implies that being female may be associated with an increased likelihood of coughing blood. Yet again this effect lacks significance.

Regarding air pollution, The positive coefficient with a t value of 17.8373 indicates a correlation between air pollution and the probability of experiencing levels of coughing blood.

```
> summary(model1)
Call:
polr(formula = df$Coughing.of.Blood ~ Age + Gender + Air.Pollution,
    data = df, Hess = TRUE)

Coefficients:
                Value Std. Error t value
Age            0.002929   0.004871  0.6013
Gender2       -0.014272   0.123702 -0.1154
Air.Pollution  0.600718   0.033678 17.8373

Intercepts:
    Value   Std. Error t value
1|2 -0.6944  0.2481    -2.7989
2|3  0.5693  0.2327     2.4464
3|4  1.5550  0.2331     6.6709
4|5  2.4201  0.2405    10.0638
5|6  2.7296  0.2451    11.1379
6|7  3.0526  0.2500    12.2085
7|8  4.3666  0.2687    16.2524
8|9  5.7726  0.2969    19.4436

Residual Deviance: 3775.862
AIC: 3797.862
```

Therefore, we will be accepting the null hypothesis given the findings.

4. Is there a connection between the level of air pollution exposure and the intensity of coughing up blood in patients?

We can use Spearman's rank correlation to assess the relationship between these factors. Spearman's correlation is a technique that doesn't rely on a connection between the variables, which makes it suitable for analyzing data.

The coefficient of Spearman's rank correlation, 0.552 indicates that there is a link between air pollution levels and the severity of coughing up blood. This suggests that as individuals are exposed to levels of air pollution the seriousness of coughing up blood tends to rise.

Null Hypothesis: There is no association between the level of air pollution and the intensity of coughing.

Alternative Hypothesis:There is an association between the level of air pollution and the intensity of coughing.

```
df$Air_Pollution <- factor(df$Air.Pollution, ordered = TRUE)
df$Coughing_of_Blood <- factor(df$Coughing_of_Blood, ordered = TRUE)

# Calculate the Spearman's rank correlation coefficient
correlation_result <- cor(as.numeric(df$Air_Pollution), as.numeric(df$Coughing_of_Blood), method = "spearman", use = "complete.obs")

# Print the Spearman's rank correlation coefficient
print(correlation_result)
```

```
> print(correlation_result)
[1] 0.5523603
```

From the above, we can reject the null hypothesis.

V.CONCLUSION

The study shows that being exposed to air pollution increases the risk of experiencing outcomes in lung cancer such as coughing up blood. Taking measures to control factors can greatly improve the prognosis. It is crucial to conduct research on how pollution, genetics, occupation and diet interact with each other in order to develop prevention and treatment methods for this illness.

VI.REFERENCES

[1] https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction