Exploratory Data Analysis of Lung Cancer Dataset

Mayank Nagar

December 4, 2023

Introduction

Lung cancer remains a global health concern. While numerous studies have been done on the association between smoking and lung cancer, the role of other lifestyle factors, such as fatigue, chest pain, and alcohol consumption, in lung cancer risk remains less explored. This study aims to investigate the relation between these variables in lung cancer dataset.

Our dataset includes information on allergies, wheezing, present of fatigue, and present of chest pain, and alcohol consumption and many more symptoms among individuals with and without lung cancer. Through rigorous statistical analysis and graphs, we want to explain the potential relationships between these symptoms/lifestyle factors and lung cancer risk.

Conducting this kind of study helps people learn about their chances of getting cancer without spending a lot of money. It also enables them to make better decisions based on their individual cancer risk. This knowledge can lead to taking actions like getting regular check-ups and making lifestyle changes that can keep them healthier.

Data Set Description

The data is collected from the Kaggle and published in the consequence of the study, "Dritsas, Elias, and Maria Trigka. 2022."Lung Cancer Risk Prediction with Machine Learning Models" Big Data and Cognitive Computing 6, no. 4: 139. https://doi.org/10.3390/bdcc6040139". In this study authors purposed machine learning methodologies to predict lung cancer based on most common habits and symptoms/signs as input features to the models.

Checking how many rows are duplicated

```
sum(duplicated(df))
```

```
## [1] 33
```

There are 33 individual for which has all same attributes.

```
dim(df)
```

```
## [1] 309  16
```

The dataset has 309 rows and 16 columns and contains following columns.

```
knitr::kable(colnames(df))
```

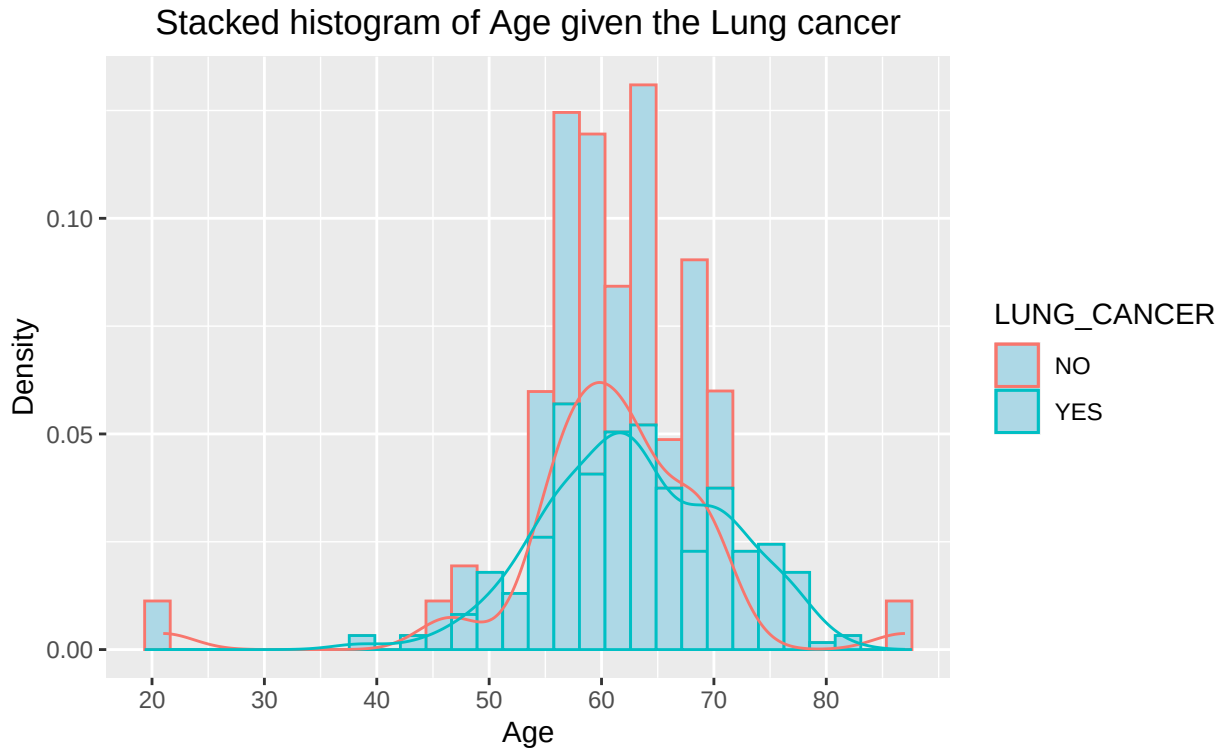
x
GENDER
AGE
SMOKING
YELLOW_FINGERS
ANXIETY
PEER_PRESSURE
CHRONIC.DISEASE
FATIGUE
ALLERGY
WHEEZING
ALCOHOL.CONSUMING
COUGHING
SHORTNESS.OF.BREATH
SWALLOWING.DIFFICULTY
CHEST.PAIN
LUNG_CANCER

Two columns have categorical values namely `GENDER` and `LUNG_CANCER` and rest 14 have numerical values.

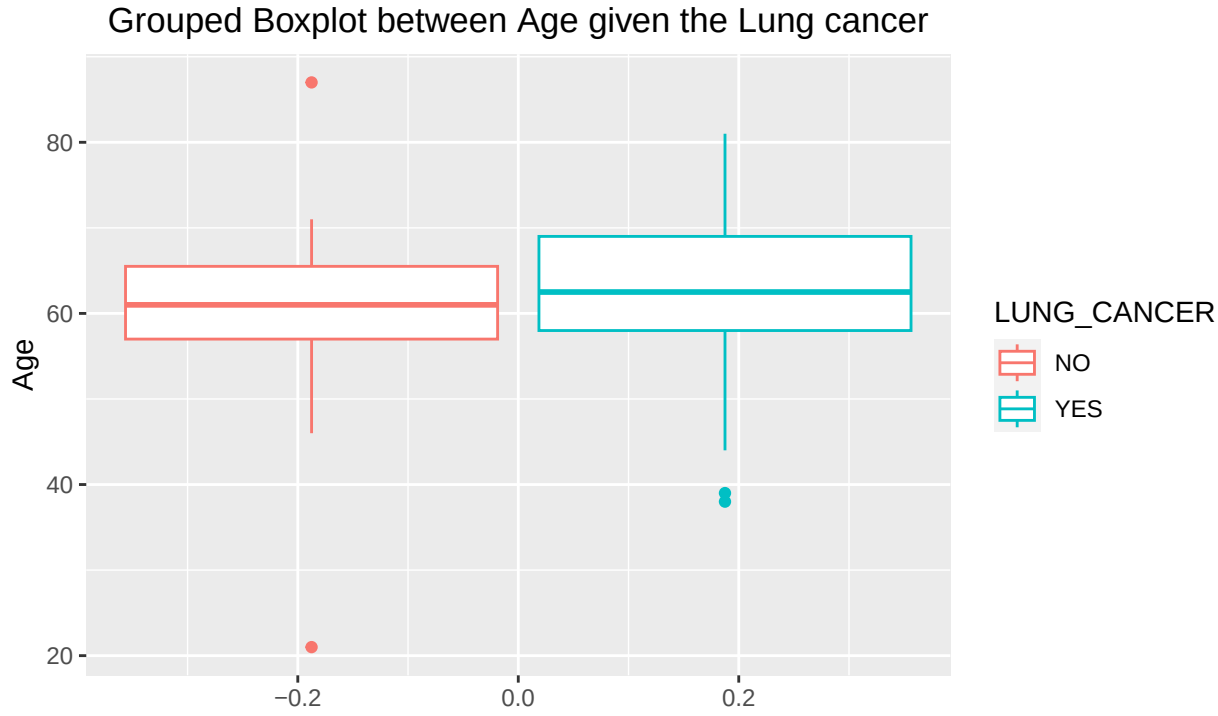
A brief detail about the features in the dataset:

1. GENDER column contains **M** for Male and **F** for Female
2. AGE column contains the age of the participant and its a numerical value
3. SMOKING columns contains 1 for non smoker and 2 for smoker participant
4. YELLOW_FINGERS column contains 1 for the person who doesn't have yellow fingers and 2 for the person who has yellow fingers
5. ANXIETY column contains 1 for the person who does not have anxiety and 2 for the person who has anxiety
6. PEER_PRESSURE column contains 1 for the person who is not facing peer pressure and 2 for the function who is facing peer pressure
7. CHRONIC.DISEASE column contains 2 for the person who has some chronic disease otherwise 1.
8. FATIGUE column contains 2 for the person who feels usually tiredness or lack of energy otherwise 1.
9. ALLERGY column contains 1 for the person who does not have any kind of allergies otherwise 2.
10. WHEEZING column contains 2 for the person who feels wheezing while breathing otherwise 1.
11. ALCOHOL.CONSUMING column contains 2 for the person who consumes alcohol otherwise 1.
12. COUGHING column contains 2 for the person who has problem of coughing otherwise 1.
13. SHORTNESS.OF.BREATH column contains 2 for the person who has the problem of shortness of breath otherwise 1.
14. SWALLOWING.DIFFICULTY column contains 2 for the person who has difficulty swallowing otherwise 1.
15. CHEST.PAIN column contains 2 for the person who has chest pain otherwise 1.
16. LUNG_CANCER column contains **YES** for the person who has the lung cancer otherwise **NO**.

Exploratory data analysis



Based on the stacked histogram, we can observe that the most frequent age range among individuals falls between 50 and 75 years, with the youngest individual being 21 years old and the oldest being 87 years old.



On the x-axis `F` stand for `Female` and `M` is for `Male`. There are some outliers present in dataset. The lung cancer is present comparatively in higher age group. Female patients with no cancer are comparatively belong to lower age group.

To identify which features are significant contributors to lung cancer, information gain is computed for each feature with respect to LUNG_CANCER column. Information gain measure the entropy loss of the dataset if we split the dataset based on the feature.

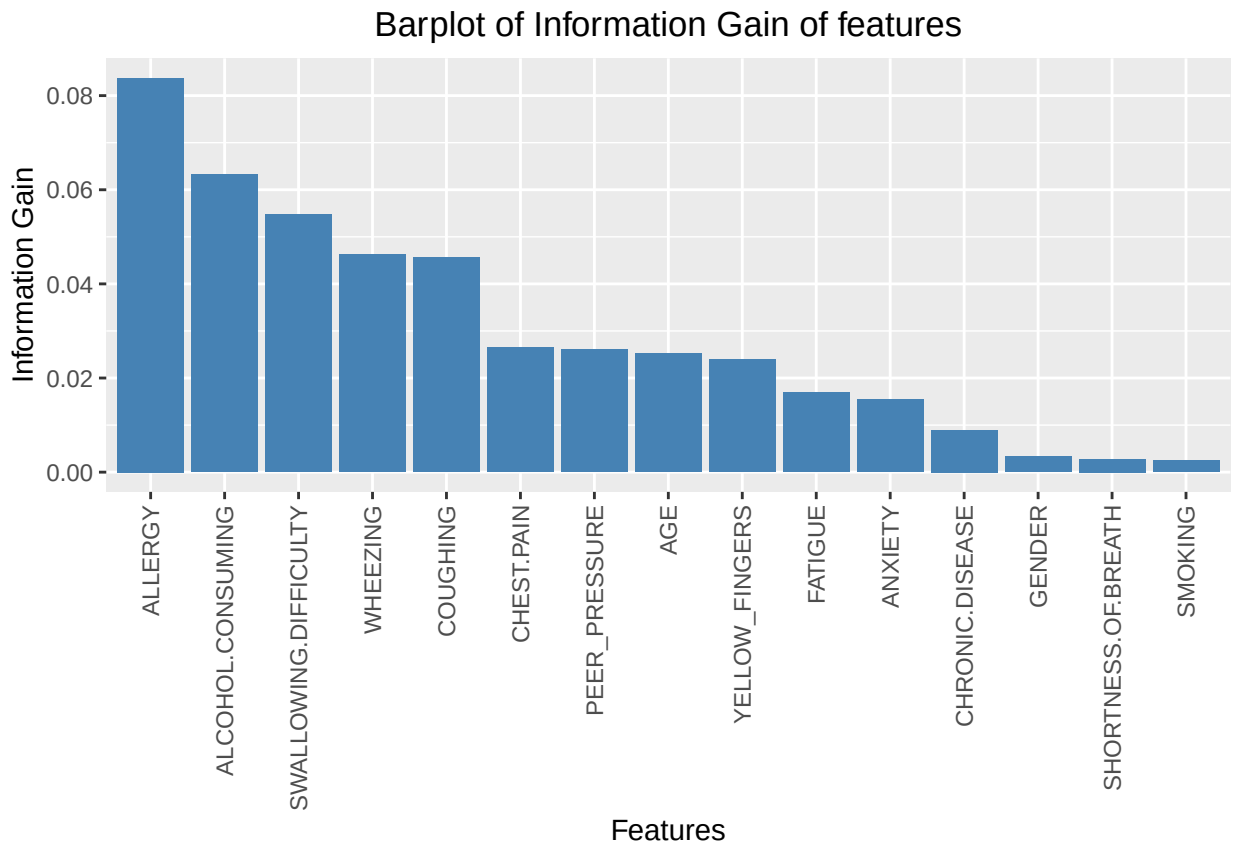
$$I.G.(S,X) = \frac{1}{Entropy(X)}\left[Entropy(S) - \sum_{v \in X} \frac{|S_v|}{|S|} Entropy(S_v)\right]$$

Where $I.G.(S,X)$ is information gain of the feature X in the dataset S. $|S|$ is number of elements in S. The formula for entropy is

$$Entropy(X) = \sum_{i=class(X)} -P_i \log_2(P_i)$$

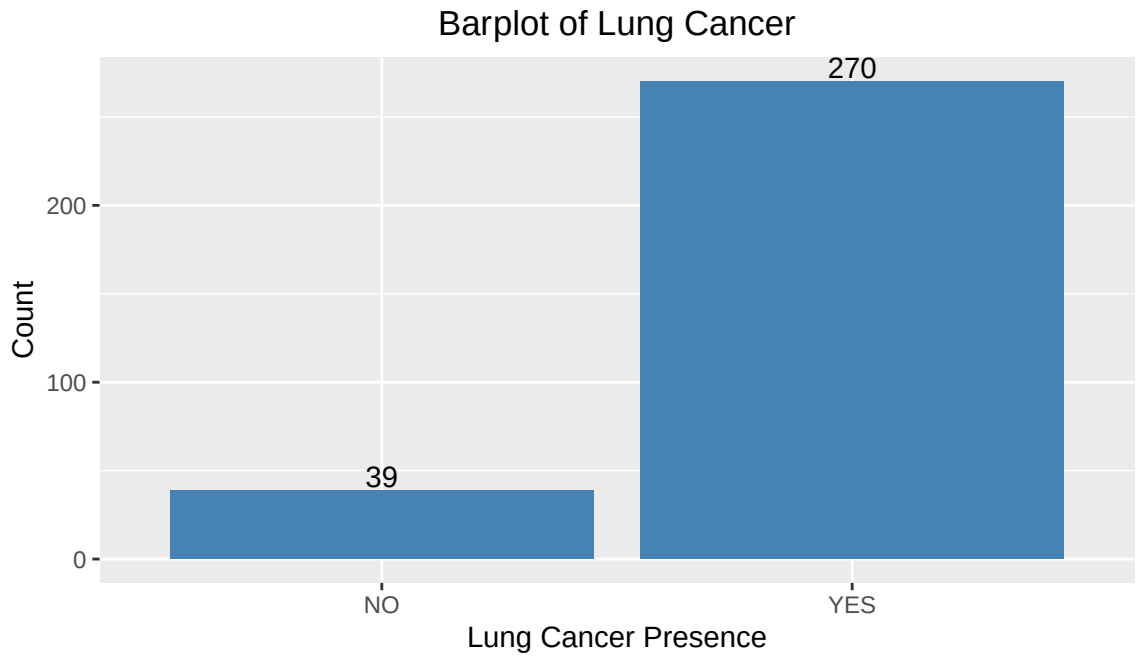
Where P_i is the observable probability of the class i in feature X.

We get the following results:



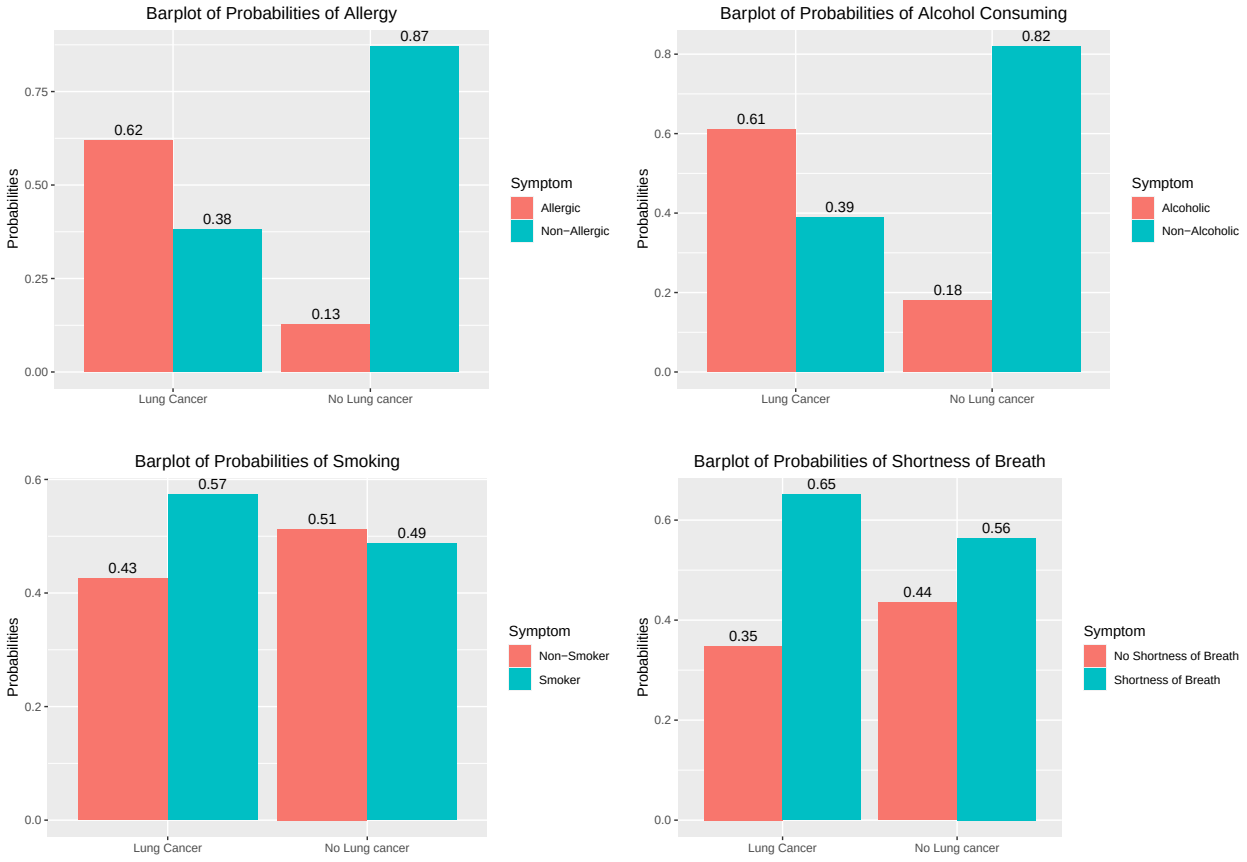
From the graph, it is clear that the variables `Allergy` and `Alcohol Consuming` provide significantly higher information gain in comparison to `Smoking`. This finding is surprising.

To verify these findings, we will create stacked barplot of individual feature with respect to lung cancer. However, before preceeding, examine the data skewness respect to `Lung Cancer`.



We observe a significant skew in the data, consisting of 39 rows of participants without lung cancer and 270 rows of participants with lung cancer. Consequently, creating a stacked bar plot of features respect to lung cancer may not give meaningful insights.

Hence, a bar plot displaying the ratio of each feature is generated given the lung cancer status.



From the graphs, followings can be concluded:

- Among individuals without lung cancer, 87% have no allergies, whereas among those with lung cancer, 62% have allergies.
- Among individuals without lung cancer, 82% are non-alcoholic, whereas among those with lung cancer, 61% are alcoholic.
- Among individuals without lung cancer, almost half are non smoker, whereas among those with lung cancer, 57% are smoker.
- Among individuals without lung cancer, 56% have shortness of breath, whereas among those with lung cancer, 65% have shortness of breath.

Therefore, `Allergy` and `Alcohol` emerge as a more prominent features associated with lung cancer compared to `Smoking` and `Shortness of Breath`.

Result

Based on the exploratory data analysis, we can draw the following conclusions:

- Most participants fall within the age group of 50 to 70 years.
- Lung cancer is more common in older age groups, while female patients without cancer tend to belong to relatively younger age groups.
- The statistical measure "Information Gain" yields surprising insights into the dataset. Features like `Allergy`, `Alcohol Consuming`, `Swallowing Difficulty`, `Wheezing`, and `Coughing` appear to be more significant for predicting Lung Cancer compared to features like `Smoking`, `Shortness of Breath`, and `Chest Pain`.
- The dataset shows an imbalance concerning the Lung Cancer feature, with only 39 participants without lung cancer and 270 participants with lung cancer.
- Among individuals without lung cancer, 87% have no allergies, while among those with lung cancer, 62% have allergies. Regarding alcohol consumption, 82% of non-alcoholics are in the non-lung cancer group, compared to 61% of alcoholics in the lung cancer group. Almost half of non-smokers belong to the non-lung cancer group, while 57% of smokers are in the lung cancer group. Furthermore, 56% of individuals without lung cancer experience shortness of breath, whereas 65% of those with lung cancer report the same symptom.

Conclusion

This study provides evidence of some common misconceptions about the leading causes of lung cancer. Smoking and chest pain are not the primary factors contributing to lung cancer. Instead, the findings emphasize the significance of other variables such as allergies and alcohol consumption in determining an individual's risk of developing lung cancer.

These findings shows how important it is to live a healthier life and making positive choices in daily habits. Maintaining a lifestyle that reduces risk factors like smoking and excessive alcohol consumption can significantly impact one's overall health and well-being. Recognizing the early signs and symptoms associated with lung cancer, such as shortness of breath or persistent coughing, should prompt individuals to seek medical assistance promptly.

Regular checkups and screenings provide opportunities for early detection, which can be helpful in improving health outcomes. By paying attention to these findings and making informed choices, individuals can take meaningful steps toward reducing their risk of lung cancer and promoting their overall health and longevity.