Lung Cancer Prediction

Introduction to Dataset:

We will be using a lung cancer prediction dataset for the duration of the research. This information was gathered from a survey to determine how air pollution affects the incidence of lung cancer. Data on the age, gender, exposure to air pollution, alcohol use, dust allergy, occupational risks, genetic risk, chronic lung disease, balanced diet, obesity, smoking, and passive smoking, as well as symptoms like chest pain, fatigue, weight loss, shortness of breath, wheezing, difficulty swallowing, clubbing of fingernails, and snoring, are all included in this dataset.

With 1.59 million fatalities from the disease in 2018, lung cancer was the world's most common cancer-related mortality cause. Smoking is the primary cause of lung cancer, although air pollution exposure is also a risk factor. According to a recent study, even among non-smokers, air pollution may raise the risk of lung cancer.

The investigation, which was reported in the journal Nature Medicine, examined information from more than 462,000 Chinese subjects who had been monitored for an average of six years. People who lived in high-pollution regions and people who lived in low-pollution areas were split into two groups among the participants.

The individuals in the high-pollution group had a higher risk of lung cancer than those in the low-pollution group, according to the researchers. Additionally, they discovered that the risk rose with age and was greater in nonsmokers than in smokers.

This study raises the possibility that there may be a connection between air pollution and lung cancer, even though it does not establish it. Additional investigation is required to verify these results and ascertain the impact that various types and concentrations of air pollution may have on the chance of developing lung cancer.

Source: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link?resource=download

Outcome of EDA:

For the analysis, I have used 5 variables:

1. Air_Pollution
2. Alcohol_Use
3. Genetic_Risk
4. Dust_Allergy
5. Chronic_Lung_Disease

- The data set has around 26 columns and 1000 rows.
- This data set has no null values and no wrong values in the columns
- Most of the columns are categorical and they are in the given range of the data.
- Most of the columns (Categorical) are in the range of 1 to 8 or 1 to 7
- Each column in the dataset is the affect that causes the cancer
- We can observe that the level of cancer is high when the range is higher for the columns which depicts the causes for the lung cancer.
- The variables don’t contain any outliers which means that all the values are inside the range

- The most common level of Air pollution and Chronic Lung Disease which is seen in most of the patients is level 6
- The most common level of Dust Allergy and Genetic Risk which is seen in most of the patients is level 7
- When we tried to analyze the PMF taking a scenario where comparing the number of patients in the lower levels to the higher levels it is observed that people in contact with higher level of Air pollution are diagnosed with the disease.
- The cumulative distribution graph of the air pollution has also showed the same outcome. The most common level of air pollution according to the CDF will be 6 which means that most of the patients in the given data have been exposed to a level 6 air pollution. We can observe that there is just 25% of the people who are affected by the air pollution under the level of 1-3 but we can observe that more than 75% has the exposure to the higher levels of air pollution.
- When we compared the correlations between the Alcohol use and Genetic risk it's Pearson Correlation is more than the spearman's Correlation.
- The logistic regression model accuracy is 76.95%. An R-squared score of 0.7695 suggests that the independent variables in the model can account for around 76.95% of the variability in the dependent variable. This number ranges from 0 to 1, with 0 indicating that the model explains no variability and 1 indicating that it fully accounts for all variability.

What do you feel was missed during the analysis?

While I was working with the dataset, I felt that this particular data set have the columns which have their own unique meaning and I was hard to correlate them and analyze them. I took time to find the variables and understand how they affect the cause of the lung cancer. So, at the initial stage of the analysis it took time to understand each categorical values and how they affect the total outcome of the data. But then when we are trying to correlate the variables and understand how they are scattered it was easy to understand the relation between the data and the variables and analyze them and apply the methods on them.

Were there any variables you felt could have helped in the analysis?

Yes, when I was trying to analyze the correlation and regression analysis, I found that most the variables I have chosen are not correlated and the regression values are very low and then when I tried to find which other variables have more correlation, I found that the 'OccuPational _Hazards' is mostly correlated to the other variables and it has provided a very good accuracy.

Were there any assumptions made you felt were incorrect?

When I was doing the initial regression analysis, I have considered the variables 'Air_pollution' and 'chronic_Lung_disease' but found that the accuracy level was very low then I had to analyze the correlation between all the variables and then find the variables which can provide the accuracy percentage above 75%.

What challenges did you face, what did you not fully understand?

The challenges that I faced during the analysis is understanding the correlations between the variables and chosing the right variables to analyze. The other challenge was framing the statistical question and hypothesis and understanding the hypothesis testing.