Lung Cancer Report

Matteo Garrò

July 21, 2024

Abstract

Lung cancer is the leading cause of cancer death worldwide, accounting for 1.59 million deaths in 2018. The majority of lung cancer cases are attributed to smoking, but exposure to air pollution is also a risk factor. A new study has found that air pollution may be linked to an increased risk of lung cancer, even in nonsmokers.

1 Introduction

This dataset contains information on patients with lung cancer, including their age, gender, air pollution exposure, 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 and snoring

2 Problem

Lung cancer is the leading cause of cancer death worldwide, accounting for 1.59 million deaths in 2018. The majority of lung cancer cases are attributed to smoking, but exposure to air pollution is also a risk factor. A new study has found that air pollution may be linked to an increased risk of lung cancer, even in nonsmokers.

The study, which was published in the journal Nature Medicine, looked at data from over 462,000 people in China who were followed for an average of six years. The participants were divided into two groups: those who lived in areas with high levels of air pollution and those who lived in areas with low levels of air pollution.

The researchers found that the people in the high-pollution group were more likely to develop lung cancer than those in the low-pollution group. They also found that the risk was higher in nonsmokers than smokers, and that the risk increased with age.

While this study does not prove that air pollution causes lung cancer, it does suggest that there may be a link between the two. More research is needed to confirm these findings and to determine what effect different types and levels of air pollution may have on lung cancer risk

2.1 Methodology

In the first part of the work an EDA has been performed in order to check for missing values, correlation between features and some transformations to the data have been performed. This made easier perform the subsequent analysis.
No missing values have been found, correlation between feature have revealed some high correlation factors between some of the features but it has been decided to not drop any of them because this solution seemed to potentially eliminate too much information from the data.
All the images for the EDA are available in the Python notebook. The second part involves classification techniques applied to the dataset.

2.2 Results

Various classification techniques have been applied to the dataset. The results can be found in a table inside the .ipynb file. The best performance has been attained by the Random Forest Classifier. For this classifier the most important feature has been **Coughing of Blood**

2.3 Final Remarks

This side project has been an interesting way to further explore the methodologies and techniques in a different setting than the one I'm used. For this reason the actual accuracy of the results obtained was not the main objective of this work. It would be interesting to further analyze this data with experts in the Oncology field in order to produce more significant results.