To: the Healthcare Community

From: Jessica Wang (jjw255), Victoria Zhang (vz36), and Jock Li (jl2698)

Subject: Leveraging Data Science to Predict the Likelihood of a Patient Developing Lung Cancer

Date: March 13th, 2024

Course code: ORIE 4741

Github link: https://github.com/jewang25/orie4741-final

Research Question:
Can we predict the likelihood of a patient developing lung cancer?

Problem Statement:
The problem is important because lung cancer stands as the leading cause of cancer-related mortality in the United States. The development of a reliable predictor holds substantial significance in the medical community, offering a pragmatic approach to inform patient care. By leveraging a dependable predictor, healthcare professionals can offer timely interventions, facilitating early diagnoses and consequently improving the prognosis for individuals afflicted with lung cancer. This not only enhances the likelihood of successful treatment outcomes but also highlights the critical role of early detection in mitigating the impact of this disease.

Dataset:
The dataset we selected includes extensive information on a thousand patients that have been diagnosed with lung cancer. This information includes age, gender, genetic risk, and other personal features. By finding the correlation between the features and lung cancer, we can find which features are more significant. Then using machine learning algorithms and data analytic techniques we can use these features to effectively predict the likelihood of lung cancer. Thus, the data set will allow us to answer our research question. The link to the dataset is included below:
https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link?resource=download

Conclusion:
Utilizing a dataset to predict whether a patient has lung cancer holds significant value for the healthcare community because lung cancer remains a predominant cause of cancer-related mortality in the United States, emphasizing the urgent need for improved diagnostic tools.

Using this dataset to create a predictor would likely succeed due to the dataset containing multiple features, including the patient's age, gender, air pollution exposure, allergies, genetic risk, fingernail characteristics, and many more. By leveraging this comprehensive dataset, predictive models can identify subtle patterns and correlations that may not be readily apparent to human clinicians. This data-driven approach enables the creation of a predictive algorithm capable of potentially accurately identifying individuals at risk of lung cancer.