Predicting Lung Cancer Incidences from Air Quality and Smoking Data

Emily Wang (20776186), Eric Yuan (20677509), Dhruva Rajkumar (20776130), Emily Tou (20845748), Hojun Choi (20845574)

Identify Need

Lung cancer is a leading cause of global death. Air quality is an environmental factor that affects the development and progression of lung cancer. This project addresses the need for more understanding of the relationship between air quality and lung cancer. We aim to contribute to identifying patterns and risk factors of lung cancer. Our project can also contribute to guiding policy for reducing public exposure to pollutants and optimizing resource allocation in the healthcare system. This project benefits individuals susceptible to lung cancer, healthcare professionals and policymakers.

Approach

Our project will utilize the Air Quality-Lung Cancer Dataset from Harvard Dataverse [1] to examine the link between air pollution and lung cancer incidence. This dataset consolidates air quality information from the U.S. Environmental Protection Agency [2] with lung cancer incidence rates from the National Cancer Institute state cancer profiles [3]. If needed, we will integrate the Smoking Prevalence Dataset [4] , also sourced from Harvard Dataverse, to control for smoking variables in our analysis. This is to isolate the specific impact of air quality on lung cancer incidence, ensuring that the effect of smoking, a well-known risk factor for lung cancer, doesn't skew the results. Our integrated approach aims to enhance the accuracy of lung cancer predictions by considering a broader set of risk factors. In our preprocessing stage, we'll merge the air quality and smoking datasets using an inner join to ensure all records are complete. During the exploratory data analysis, if needed, we'll apply feature scaling, selection, or regularization, and consider outlier treatment to prepare our data for modeling.

In our model selection, we'll use supervised learning algorithms since we will be working with labeled data, where an output (lung cancer incidence rates) will be predicted by known inputs. Potential supervised learning models that will be used for this project will include, linear regression, decision trees, random forest and support vector regression. Each model will undergo hyperparameter tuning for optimal performance, ensuring accurate and reliable predictions.

Comparison and Validation

To compare methods and validate results, we will first use cross validation. We will partition the dataset into k subsets, training the model on (k minus 1) subsets and validating on the remaining subset. We will repeat this k times to ensure that each subset has been used for both training and validation. This will reduce overfitting and also assess different parts of the data on the model's performance.

The dataset will be divided randomly into training and testing sets. The model will be trained on the larger set and evaluated on the smaller set.

We will also validate the results of our model on datasets from different time frames and locations. This will evaluate the model's ability to work for diverse scenarios. This will also detect whether the results are true regardless of the original datasets' location and time frame.

We will also find additional datasets related to air quality and lung cancer incidences from different sources. We will run the model on different datasets to assess our model's transferability.

We will implement and compare the performance of multiple supervised learning methods. By implementing various models, we can validate which is the most suitable approach for predicting lung cancer incidences. We will also assess models using error metrics such as RMSE and R-squared. RMSE quantifies the average difference between predicted and actual values. R-squared evaluates the proportion of variance explained by the model. These metrics

will help us understand the model's predictive efficacy. We will select our model based on the lowest RMSE and highest R-square values. This ensures our model is accurate and interpretable.

Since overfitting can happen when the model learns the training data too well but fails to generalize. We will tune the hyper-parameters during cross-validation. This will balance complexity and generalization.

Potential Risks

A potential risk in our project arises from the temporal misalignment between the National Cancer Institute state cancer profiles and the domain-specific county-level Environmental Quality Index (EQI) dataset. The former covers the period 2010-2014, while the latter spans 2000-2005. This discrepancy may affect the synchronization of variables, potentially introducing errors and inaccuracies in our predictions. The dataset does, however, contain averages of time frames from either 2010-2014 or 2000-2005, mitigating the majority of potential errors resulting from the time difference.

Another potential risk stems from inaccuracies and missing values within the Air Quality-Lung Cancer dataset, particularly in the variables [Inter, Slp, control, treat, LocalTreat]. These variables, crucial for identifying control and treatment groups, only have data for approximately 54 out of 2602 counties. This limited coverage may lead to non-informative results, especially in understanding the type and extent of treatment specific counties have received. Without comprehensive information on the treatment specifics, the data could become skewed or inaccurate, impacting the validity of our analyses. Regular validation techniques will be employed to address missing values and ensure the robustness of our findings.

Other general potential risks may arise while working on our project such as overfitting during the hyperparameter tuning process which may occur when overly optimizing the model for the training data.

Design Thinking and Project Management

Our plan in terms of communication is to use a group chat and have weekly meetings where we can discuss our project. Prior to each meeting there will be an agenda in order to use time more effectively. We will talk about our progress on the project regularly in order to make sure no one is impeded.

Our plan to come up with ideas stems from performing individual research on our topic and then individually proposing potential models. This way we can help eliminate potential bias and consider a wider range of ideas. We plan to divide and organize tasks using a trello board where the division is based on a first come first serve. Specific tasks will be created when more design detail on our project ideas is complete. Currently, we plan on having much more broad tasks; research/ideate, prototype 1, test 1, evaluate, prototype 2, test 2 and final delivery.

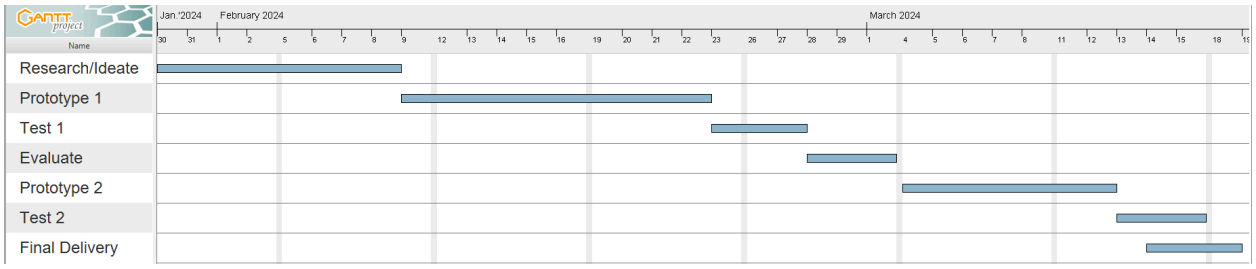

Figure 1: Gantt Chart

References

[1] "Air Quality-Lung Cancer Dataset," Harvard Dataverse, [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HMOEJO.

[2] U.S. Environmental Protection Agency, "Environmental Quality Index (EQI)," [Online]. Available: https://www.epa.gov/healthresearch/environmental-quality-index-eqi.

[3] National Cancer Institute, "State Cancer Profiles: Incidence Data," [Online]. Available: https://statecancerprofiles.cancer.gov/data-topics/incidence.html.

[4] "Smoking Prevalence Dataset," Harvard Dataverse, [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VZ21KD.