Leveraging Data Science to Predict the Severity of Lung Cancer for a Patient

Victoria Zhang (vz36), Jessica Wang (jjw255), Jock Li (jl2698)
Github Link: https://github.com/jewang25/orie4741-final
Department of Operations Research and Information Engineering, Cornell University
ORIE 4741: Learning with Big Messy Data
Professor He
May 4, 2024

Introduction

Lung cancer stands as the leading cause of cancer-related mortality in the United States and claims an immense toll of lives globally. While treatments have advanced, preventing lung cancer through early detection and risk assessment is crucial for improving outcomes. However, predicting an individual's likelihood of developing this disease has proven challenging due to the complex combination of genetic and environmental risk factors involved. To better understand how to identify those at highest risk, we analyzed a *Lung Cancer Prediction Dataset* from Kaggle that consists of information on patients with lung cancer. Ultimately, our research aims to provide healthcare providers with an empirically-grounded risk assessment tool to stratify patients for lung cancer screening and preventive interventions. By identifying the highest-risk individuals and how severe the cancer will be, we can optimize the allocation of healthcare resources and potentially save lives through early detection and risk mitigation strategies.

The Problem

Due to the deadly side effects of lung cancer, the development of reliable predictors hold substantial significance in the medical community, offering a pragmatic approach to inform patient care. By leveraging dependable predictors, 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. Therefore, we proposed a series of research questions to help us examine and ultimately solve the problem:

1. *What are the most important risk factors in predicting the severity of lung cancer and how are they related to each other?*
2. *Is it possible to predict which patients are at high risk of lung cancer based on their biometrics and if so how accurately?*

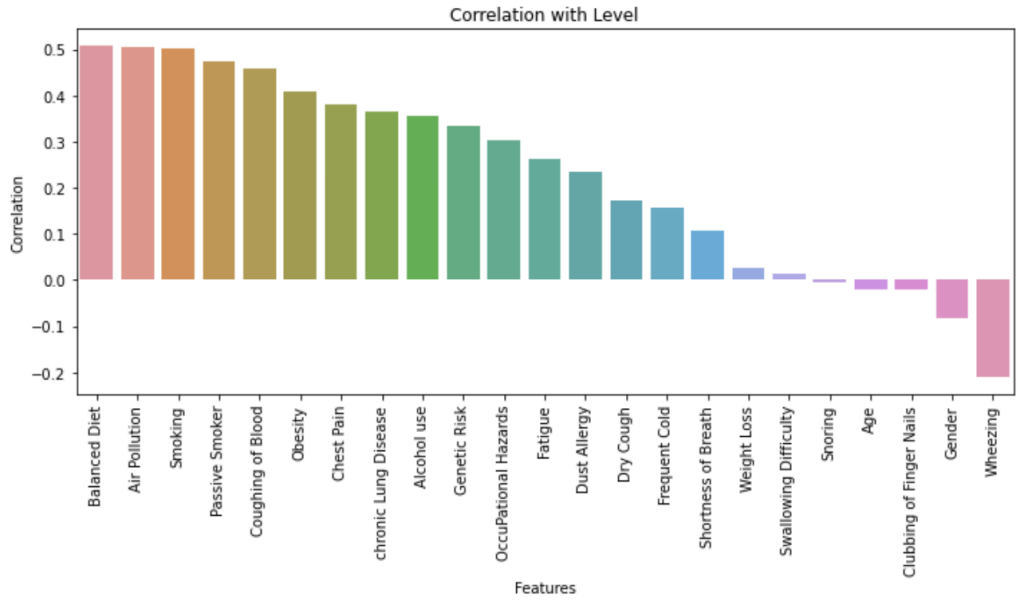
Dataset and Preprocessing

The dataset we chose contains information on patients diagnosed with lung cancer, including demographic factors, environmental exposures, lifestyle habits, medical history, and symptom presentation. The data was obtained from a large-scale study conducted in China, which followed over 462,000 participants for an average of six years. For simplicity reasons, we are only using a random sample of 1,000 participants from the total study population in our models. Participants were divided into two groups based on their residential exposure to air pollution levels, with one group living in areas with high levels of air pollution and the other in areas with low levels of air pollution. This study design allows for the investigation of the potential link between air pollution exposure and lung cancer risk, even among non-smokers. Due to the size of the data and how it was collected, we can infer that there is a lower chance that bias will play a significant factor in our analysis.

Before we can analyze the data and draw any conclusions, several preprocessing steps were undertaken to ensure data quality and integrity. We noticed that there were some instances of missing values and almost all of the data were categorical, which could pose a problem if we tried to use regression for prediction. Missing values were handled using appropriate imputation techniques, such as mean imputation for continuous variables and mode imputation for categorical variables. Categorical variables were one-hot encoded, and continuous variables were standardized to have a mean of zero and a standard deviation of one. For example, the 'Level' column consisted of low, medium and high depending on the severity of lung cancer that the patient has. We encoded low to be 0, medium to be 1, and high to be 2. Similarly, the rest of the features like air pollution, smoking, obesity, etc. were categorized on a scale from 1 to 8 where 8 is the most severe case.

Finally, we randomly split the dataset into training and testing sets, with 80% of the data used for model training and the remaining 20% reserved for model evaluation and validation. With the preprocessed dataset, we proceeded to explore the relationships between the various risk factors and lung cancer incidence, as well as to develop predictive models for estimating an individual's likelihood of developing lung cancer based on their risk profile. We started off by finding the key clinical features and seeing if or how they're correlated to each other.

Figure 1: *Correlation between All Clinical Features and Severity of Lung Cancer Level*



Our goal with this visualization is to see if there are any features that are highly correlated with a patient's level of lung cancer. Thus, based on Figure 1, the clinical features that have a statistically significant correlation with the severity of lung cancer level include *Balanced Diet, Air pollution, Smoking, Passive Smoker, Coughing of Blood, Obesity, Chest Pain, Chronic Lung Disease, Alcohol use, Genetic Risk, Occupational Hazards,* and *Dust Allergy.* We determined based on the context and the size of our dataset that correlations with an absolute value greater

than 0.3 are considered significant and should be explored more. Since these are all positive correlations, we claim that as the value of the clinical feature increases, the severity of lung cancer level tends to increase as well. The features with the highest correlation are *Balanced Diet, Air pollution,* and *Smoking*. Higher levels of air pollution and smoking intuitively could cause someone to develop more severe lung cancer, however having a balanced diet also had a positive correlation. We will further analyze these significant variables by seeing how they're correlated to each other.

Appendix A contains what we consider to be the most significant medical features listed in the dataset as well as their relationships with each other. To determine how correlated the variables are to each other, we want to look at the intersections between each pair of features, where the higher the value is, the more correlated they are. This value is determined by taking the covariance of each datapoint and weighting that against the variance of the factor. As shown by the scale on the left hand side of the visualization, the more red the boxes are, the higher the correlation. Immediately we can see that *Genetic Risk* and *Occupational Hazards* are the most correlated with a value of 0.89, followed by *Occupational Hazards* and *Alcohol Use*, *Genetic Risk* and *Alcohol Use*, and finally *Chronic Lung Disease* and *Occupational Hazards*. These combinations, along with several others shown in the heatmap may play a crucial role in predicting the severity of lung cancer.

Predicting the Severity of Lung Cancer

The Three Techniques

Predicting the level of severity of lung cancer given various symptoms is a multiclassification problem. There are various models that can be used to classify data into multiple classes. We decided to focus on three models: support vector machines (SVM), logistic regression, and trees. Through the results from these models, we will be able to determine which type of model is most effective for our dataset and which features are most important for predicting the severity of lung cancer.

SVM

SVM can be used to classify data by determining the optimal decision boundary to separate the data into different classes. Since SVM is a linear model that can be used on nonlinear datasets, we can use it to try to classify the data into three classes based on the severity of lung cancer (low, medium, and high). Although SVM works better and is easier to visualize on smaller datasets with less features, it can be used to see if there is a nonlinear separator for the data that we can use to predict the severity of lung cancer.

We cannot easily visualize the separator for the data when there are many features included in the SVM model. However, we can use accuracy and f1 scores to determine how well the model worked on the data and rank the features by importance. As our dataset is nonlinear, we ran the

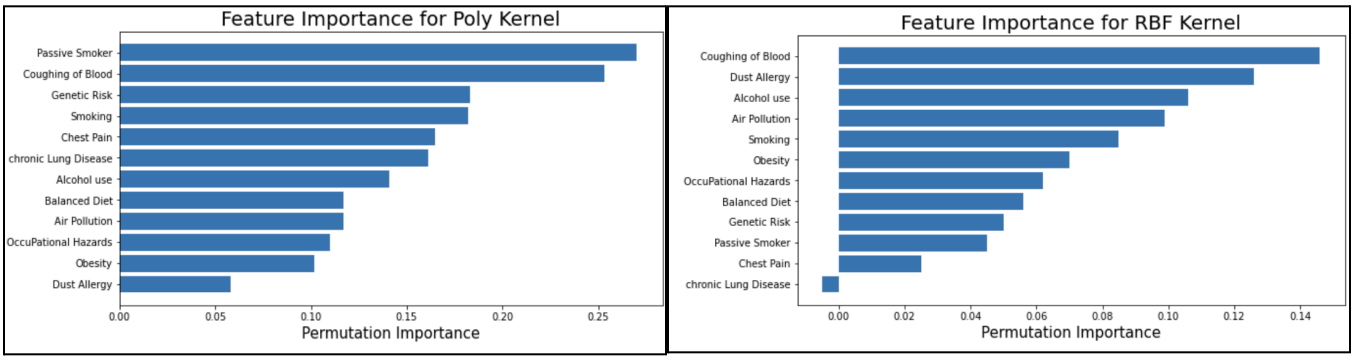
radial basis function and polynomial kernel in SVM to see which one would work the best. The accuracy score measures the number of correct predictions made by the model while the f1 score measures the precision of the model overall.

Table 1: *SVM Accuracy and Precision*

Kernels	Accuracy Score	F1 Score
Polynomial (poly)	100%	100%
Radial Basis Function (rbf)	99%	99%

Permutation importance reflects the change in model performance when values of a feature are changed and the other features are kept the same. We can use the permutation importance function from sklearn to find which features are most important in the model in determining how severe a patient's lung cancer is. When looking at the graphs below, it appears that coughing of blood, smoking, and alcohol use are features that are significant in helping classify the severity of one's lung cancer. Intuitively, these activities or symptoms seem to decrease one's lung health which increases the likelihood of one's lung cancer being more severe.

Figures 3 and 4: *Feature Importance Bar Graph*



Logistic Regression

Based on our previous analysis of the clinical features that showed strong correlations, we can conduct a more complex predictive analysis to assess the severity of a patient's lung cancer. To do this, we will use a machine learning algorithm called logistic regression to predict the level of a patient's lung cancer. Logistic regression is a statistical model and data analysis technique that explains the relationship between a binary variable or in this case multi-label classification. As we mentioned earlier, each patient is labeled with 0, 1, or 2 corresponding to the severity of lung cancer (low, medium, and high). This will ultimately allow us to figure out which of the selected clinical features in our dataset are the best predictors for severity of lung cancer.

Table 2: *Summary of the Logistic Regression Output for Severity Level Medium*

Variables	**β**	**S.E**	**p-value**	**95% C.I.**
Air Pollution	1.4486	0.197	< 0.001	[1.062, 1.836]
Alcohol Use	-3.0298	0.272	< 0.001	[-3.563, -2.496]
Dust Allergy	1.2314	0.168	< 0.001	[0.901, 1.561]
Occupational Hazards	1.8474	0.257	< 0.001	[1.344, 2.351]
Genetic Risk	0.0902	0.265	0.733	[-0.429, 0.609]
Chronic Lung Disease	0.0867	0.152	0.569	[-0.211, 0.385]
Balanced Diet	-0.5171	0.165	0.002	[-0.841, -0.193]
Obesity	-1.2285	0.193	< 0.001	[-1.607, -0.859]
Smoking	1.1775	0.146	< 0.001	[-0.891, -0.193]
Passive Smoker	-0.8088	0.143	< 0.001	[-1.090, -0.528]
Chest Pain	-0.4402	0.150	0.003	[-0.733, -0.147]
Coughing of Blood	-0.0230	0.147	0.876	[-0.312, 0.266]

The table above includes the coefficients, standard errors, p-values, and confidence intervals of the features for the label *Level = 1* or when the severity of lung cancer is medium. The full results of logistic regression, which includes the results for *Level = 2* (High severity) and overall R-squared value, is shown in the appendix section. Based on these results, we can determine which features are significant predictors and how well the model performed.

Since p-values of 0.05 or less are considered statistically significant, we can observe that for *Level = 1* (medium severity), significant predictors include *Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazards, Balanced Diet, Obesity, Smoking, Passive Smoker,* and *Chest Pain.* For *Level = 2* (high severity), significant predictors include *Air Pollution, Alcohol Use, Dust Allergy, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity,* and *Passive Smoker.* Notice that the predictor *Occupational Hazards* was significant when predicting *Level =1* but not for *Level = 2*. Similarly, *Chronic Lung Disease* was significant for *Level = 2*, but not for *Level = 1*. This suggests that between the different severities of lung cancer, certain predictors are more likely to be associated to a particular level but not to others. Moreover, features like *Genetic Risk* and *Coughing of Blood* are not likely to be predictors for either level.

Having a high and positive coefficient (β) indicates a strong positive association between the predictor variable and the outcome variable. Based on our model, features like *Air Pollution, Dust Allergy, Occupational Hazards,* and *Smoking* all have large, positive coefficients, meaning

that higher levels of these predictors are associated with an increased likelihood of having medium severity lung cancer. The same goes for high severity lung cancer with an emphasis on *Dust Allergy*. Note that *Alcohol Use* has a large, negative coefficient, which implies that higher alcohol use is associated with a decreased likelihood of being in medium/high severity of lung cancer. These results make sense because the significant features are all associated with harming the lungs directly, while the rest have a negative effect on the overall body.

Trees

Decision trees are in general a popular supervised learning algorithm used for classification. The algorithm partitions the feature space into regions, with each partition corresponding to a specific outcome or prediction. We decided to use this technique because they are easy to interpret, and can handle both numeric and categorial data. Therefore, in our lung cancer classification dataset, decision trees would be advantageous due to having both numerical and categorical data. Decision trees have a tendency to create overly complex models that can memorize the training data which would lead to poor generalization. We initially generated the following decision tree after utilizing the sklearn decision tree classifier. The accuracy of the decision tree was 1.0, indicating that the model was overfitting our data. Looking at Table 3 below, we can see that the most important features for the decision tree were coughing of blood, air pollution, and obesity.

Table 3: *Important Features Based on Decision Tree*

Feature	Importance
Air Pollution	0.07326541286674572
Alcohol Use	0.17211279712244454
Dust Allergy	0.0
Occupational Hazards	0.07133166776222803
Genetic Risk	0.039966993652745884
Chronic Lung Disease	0.021141865919190424
Balanced Diet	0.03151200553408419
Obesity	0.0837201081495942
Smoking	0.0
Passive Smoker	0.0
Chest Pain	0.0
Coughing of Blood	0.506949148992967

Boosting is necessary in our context to enhance the performance of our decision trees in solving the problem of overfitting our data. We ended up utilizing the AdaBoost algorithm to address this issue by sequentially training multiple weak learnings on different subsets of data, with each subsequent learner focusing on the mistakes made by the previous learners. Boosting would then hopefully produce a strong ensemble model that generalizes well to new data, which would improve the accuracy as well as robustness of the classification of lung cancer levels. The accuracy ended up decreasing from 1.00 to 0.90, indicating a more generalized model for predicting the level of severity of lung cancer.

Table 3: *Accuracy of Tree Models*

Model	**Accuracy Score**
Decision Tree	100%
Decision Tree with Boosting	90.5%

Results and Conclusions

To predict the severity and extent of lung cancer a patient has based on certain biometrics and daily life exposures, we utilized SVM, logistic regression, and decision trees. By applying SVM on our dataset, we discovered that *Coughing of Blood*, *Smoking*, and *Alcohol Use* were prominent features to classify the severity of one's lung cancer. The accuracy score and F1 score for the Radial Based Function was 99%. In logistic regression, the pseudo R-squared value of 0.5444 suggested that the model explains a moderate amount of the variability in the severity levels. Moreover, the very low log-likelihood ratio (LLR) p-value indicates that the model as a whole is statistically significant compared to a null model. Therefore, we can conclude with some degree of confidence that predictors *Air Pollution, Dust Allergy, Occupational Hazards,* and *Smoking* have an important role in predicting the severity of lung cancer relative to the other features. After utilizing decision trees, there was significant overfitting when utilizing the features with high correlation from Figure 1. Therefore, we used a decision tree with boosting, and got an accuracy of 90.5% which allowed for more generalized data and a more robust model.

To compare the models, we selected these models SVM, logistic regression, and trees because the dataset had mainly categorical data. The variation in the most important features across different models can be attributed to the inherent differences in how each algorithm approaches the task of feature selection and model fitting. In comparing the performance of these algorithms, SVM demonstrated accuracy and robustness, particularly in capturing nuanced relationships among features, as evidenced by its high accuracy score of 99%. However, logistic regression provided valuable insights into the explanatory power of the model, offering a moderate but statistically significant explanation of the variability in lung cancer severity. Decision trees alone faced challenges with overfitting due to correlated features, highlighting the importance of employing boosting techniques to enhance generalization. Ultimately, the decision tree with

boosting approach achieved a commendable accuracy of 90.5%, striking a balance between model complexity and predictive performance. While each algorithm exhibited strengths and weaknesses, the utilization of decision trees with boosting emerged as the most effective strategy for achieving both accuracy and generalization in predicting lung cancer severity from the dataset's categorical features.

Final Recommendations and Next Steps

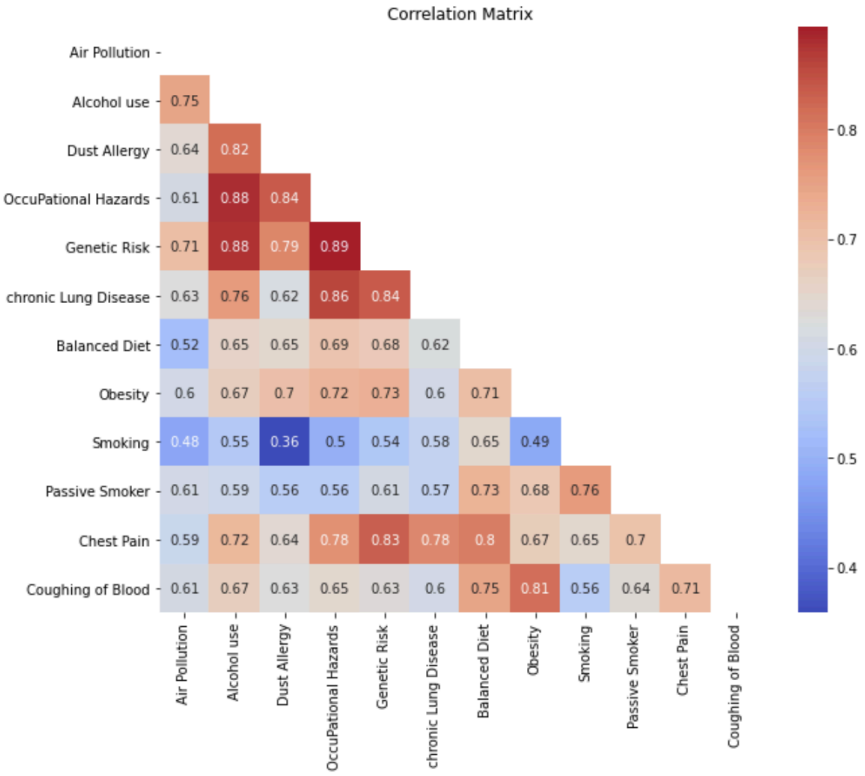
As our models were trained on participants from a study done in China, its predictions could be less accurate for people from other countries. Our dataset also only includes patients in hospitals which could skew our results towards individuals with higher socioeconomic statuses. This could highlight factors that are prevalent in higher socioeconomic classes, enhancing existing societal inequality just like a “weapon of math destruction.” If we were to improve the fairness of our machine learning models, we would try to remove certain features such as genetic risk or group them based on race and socioeconomic status to eliminate potential bias. One limitation of our models is that the data is only collected from patients who give consent due to privacy concerns which limits the sample the models are trained on. The models also may not generalize well because many of the features are categorical and the model is not trained on a very diverse population. In the future, we would like to extend our models to analyze important features more closely and help determine what treatments are best for different levels of lung cancer.

Contributions

Jessica worked on building the SVM model and figuring out what the next steps for this project should be.Victoria worked on utilizing a Decision Tree as well as a Decision Tree with Adaboost on our dataset and writing the results and conclusions for the project. Jock worked on the introduction, data analysis/processing, and building the logistic regression.

Appendix

Appendix A: *Key Clinical Features Heatmap*



Appendix B: *Full Logistic Regression Results with Selected Clinical Features*

MNLogit Regression Results

Dep. Variable:	Level_encoded	No. Observations:	1000
Model:	MNLogit	Df Residuals:	976
Method:	MLE	Df Model:	22
Date:	Tue, 07 May 2024	Pseudo R-squ.:	0.5444
Time:	13:03:37	Log-Likelihood:	-499.19
converged:	True	LL-Null:	-1095.7
Covariance Type:	nonrobust	LLR p-value:	1.359e-238

Level_encoded=1	coef	std err	z	P>|z|	[0.025	0.975]
Air Pollution	1.4486	0.197	7.336	0.000	1.062	1.836
Alcohol use	-3.0298	0.272	-11.126	0.000	-3.563	-2.496
Dust Allergy	1.2314	0.168	7.312	0.000	0.901	1.561
OccuPational Hazards	1.8474	0.257	7.188	0.000	1.344	2.351
Genetic Risk	0.0902	0.265	0.341	0.733	-0.429	0.609
chronic Lung Disease	0.0867	0.152	0.570	0.569	-0.211	0.385
Balanced Diet	-0.5171	0.165	-3.129	0.002	-0.841	-0.193
Obesity	-1.2285	0.193	-6.368	0.000	-1.607	-0.850
Smoking	1.1775	0.146	8.051	0.000	0.891	1.464
Passive Smoker	-0.8088	0.143	-5.644	0.000	-1.090	-0.528
Chest Pain	-0.4402	0.150	-2.942	0.003	-0.733	-0.147
Coughing of Blood	-0.0230	0.147	-0.156	0.876	-0.312	0.266

Level_encoded=2	coef	std err	z	P>|z|	[0.025	0.975]
Air Pollution	-0.5937	0.135	-4.388	0.000	-0.859	-0.329
Alcohol use	-1.5939	0.260	-6.122	0.000	-2.104	-1.084
Dust Allergy	2.2186	0.180	12.354	0.000	1.867	2.571
OccuPational Hazards	0.1100	0.294	0.374	0.708	-0.466	0.686
Genetic Risk	0.7115	0.255	2.786	0.005	0.211	1.212
chronic Lung Disease	0.4708	0.185	2.538	0.011	0.107	0.834
Balanced Diet	-0.7370	0.132	-5.602	0.000	-0.995	-0.479
Obesity	-0.6828	0.203	-3.356	0.001	-1.082	-0.284
Smoking	0.0997	0.119	0.838	0.402	-0.133	0.333
Passive Smoker	-0.3829	0.148	-2.584	0.010	-0.673	-0.092
Chest Pain	-0.0916	0.164	-0.558	0.577	-0.413	0.230
Coughing of Blood	0.0013	0.147	0.009	0.993	-0.287	0.290

Appendix B: *Full Decision Tree*