MLPH Final Project Report - Lung Cancer Prediction Models

Farah Beche

Introduction

Globally, lung cancer is the primary cause of cancer-related deaths and stands as the third most common cancer in the United States.[2] Individuals who attain risk factors such as substance use, chronic diseases, and increased age are particularly susceptible. Lung cancer arises when cells within the lungs proliferate without control, creating a lump known as a tumor. These growths may be benign (non-cancerous) or malignant (cancerous). However, this cannot be known without being tested. Traditional diagnostic methods include physical examinations, CT scans, ultrasounds, and biopsies. Although they are effective, they present drawbacks. Biopsies, in particular, are invasive, uncomfortable, stress-inducing, and financially burdensome. Additionally, after diagnosis, survival rates vary based on numerous factors, with the type of lung cancer playing a crucial role. This refers to the specific type of cell from which the cancer originated, such as non-small cell lung cancer or small cell lung cancer. Moreover, depending on the stage of lung cancer, a majority of people diagnosed with NSCLC and SCLC between 2012 and 2018 had low 5-year relative survival rates.[3] Consequently, there's a pressing demand for the development of non-invasive, precise, and accurate diagnostic alternatives. Predictive models could significantly improve how we diagnose and treat lung cancer by potentially avoiding invasive tests like biopsies and low survival rates.

In this project, machine learning methods are applied to
assist in lung cancer prediction is based on a number of risk factors including gender, age, smoking status, etc. which makes it a powerful tool for both individuals at risk and doctors. For individuals, these methods can provide personalized risk assessments based on their lifestyle factors such as age, smoking history, and chronic conditions, leading them to make informed decisions about screening and lifestyle changes. For doctors, machine learning models can assist in early detection by analyzing vast amounts of patient data to identify patterns indicative of lung cancer, leading to earlier intervention and improved patient outcomes. A total of six different machine learning methods are selected in this project, namely Ridge, LASSO, Logistic Regression, K-nearest neighbor, Decision Tree, and Random Forest. It is hypothesized that the final test outcome would achieve 90% accuracy in detecting the most impactful risk factors for causing lung cancer.

Related Work

Researchers have established various machine-learning models for lung cancer diagnosis, treatment, and prognosis[4]. For instance, some models use sequencing data and machine-learning techniques for treatment and diagnosis[4]. However, the practicality of these techniques is limited due to the challenges and expense associated with obtaining DNA and RNA sequences necessary for analysis. Therefore, there is a need for a simpler, more cost-effective, and accurate method that focuses on analyzing risk factors.

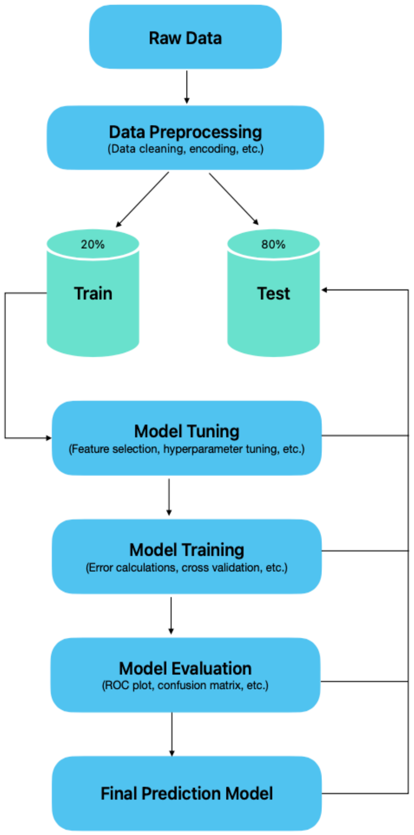
While there are numerous applications of machine learning in lung cancer classification, the dataset and models remain valuable due to its uncomplicated and accessible nature, which requires no specialized training for classification. After reviewing the report, even those new to machine learning can conduct the analysis confidently. This approach offers a straightforward yet effective way to leverage machine learning for lung cancer classification, enhancing accessibility and usability in clinical settings.

Methods

The aim of this project is to develop an accurate predictive model for lung cancer based on various risk factors, including demographic and lifestyle factors. The predictor variables included gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty, and chest pain. The response variable was the presence or absence of lung cancer. In order to find the optimal model for predicting lung cancer most accurately, various classification models were used including logistic regression, KNN, decision tree, and random forest with ridge and lasso methods being performed for feature selection. Each of these algorithms offered unique advantages and were well-suited for different aspects of lung cancer prediction.

Logistic regression provided a straightforward interpretation of coefficients, making it easy to understand the impact of each predictor variable on the outcome. However, logistic regression assumes a linear relationship between the predictor variables and is sensitive to outliers, which may not capture complex relationships between predictors and the outcome.

K-Nearest Neighbors can capture complex relationships in the data and is easy to understand. However, the choice of the number of neighbors (K) is critical, and selecting an inappropriate value can lead to poor performance. KNN also considers all features equally, so noisy features can negatively impact its performance.


Raw Data
Data Preprocessing
(Data cleaning, encoding, etc.)
20%
Train
80%
Test
Model Tuning
(Feature selection, hyperparameter tuning, etc.)
Model Training
(Error calculations, cross validation, etc.)
Model Evaluation
(ROC plot, confusion matrix, etc.)
Final Prediction Model

Decision trees capture nonlinear relationships between predictors and the outcome, which can be visualized and interpreted. However, decision trees are prone to overfitting and may not capture complex relationships as effectively as other models like random forests, especially when there are many features.

Random forests typically provide high accuracy by averaging predictions from multiple decision trees. Additionally, by averaging predictions from multiple trees, random forests mitigate the risk of overfitting present in individual decision trees. However, compared to other models, random forest models are less interpretable especially compared to individual decision trees due to the ensemble nature of the model.

By using these four models, training and testing errors will be compared to examine model performance for predicting lung cancer status. Additionally, a ROC curve plot would be produced to compare AUC values, providing a measure of the model's ability to distinguish between sensitivity and specificity. Furthermore, confusion matrices will be generated to examine the two models with the highest AUC values classification performances in more detail. The standard to determine the effectiveness and accuracy of the model would be high AUC values in a ROC plot and high accuracy rates from a confusion matrix output. A figure illustrating the overall methodology is displayed on the right.

Data and Experiment setup

The Lung Cancer DataSet is collected from an online lung cancer prediction system, designed to help individuals assess their cancer risk[1]. The dataset comprises 284 instances with 16 attributes containing information on various demographic and health-related attributes of individuals including gender, age, and substance use, along with their lung cancer status. The dataset aims to facilitate the development of predictive models for assessing lung cancer risk based on individual characteristics. We applied 80% of the original data to be training data and the rest 20% for testing.

To begin predicting the diagnosis, we first cleaned and preprocessed the dataset by checking and eliminating any repetitive data and checking for missing values. Fortunately, there were no missing values in the dataset. All the variables in the dataset were binary except for the patient's age. Among the binary predictor data, 2 represented YES (attaining the variable) and 1 represented NO (not attaining the variable). Among the binary predictor GENDER, M represented male and F represented female. Among the binary response variable, YES represented being diseased by lung cancer and NO represented not being diseased. The only encoding procedure that was done was assigning 0 to "M" and 1 to "F" for gender variable and 0 to "NO" and 1 to "YES" for the response variable.

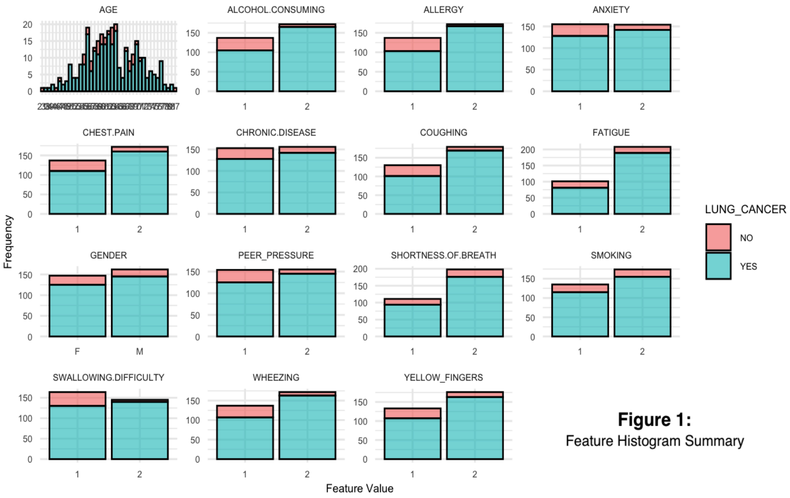


Figure 1:
Feature Histogram Summary

Figure 1, a table of summary histograms shows each feature's distribution. 309 observations were included in the data, among which 87% (270 observations) were diagnosed with lung cancer while 13% (39 observations) were lung cancer-free. In these 309 observations, 80% of them are randomly selected as the training set for model building, and the remaining 20% are included in the testing set.

Figure 2:
Correlation Matrix

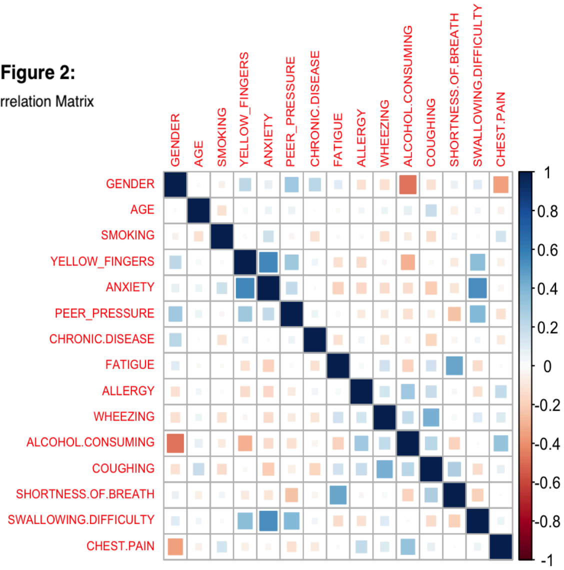


Figure 2, a correlation matrix, shows the relationships between various features in the data and helps in identifying potential multicollinearity issues. By examining the correlation matrix, we can identify highly correlated features and potentially remove less important ones. For example, alcohol consumption appears to be positively correlated with chest pain and negatively correlated with gender. However, simply looking at correlations is not sufficient for optimal feature selection.

To improve the performance and interpretability of the models, as well as reduce overfitting issues, feature selection, more specifically regularized regression techniques ridge and lasso, were performed for experimental setup to identify the most relevant features. Using 5-fold cross validation, figures 3 and 4 show the coefficient paths for ridge and lasso regressions across different values of the regularization parameter lambda. In figure 3 with larger lambda values, we can see ridge regression has heavily shrunk some features. Figure 4 provides a clearer view of where lasso starts dropping features from the model as lambda increases. By examining these coefficient paths, we can identify the most relevant features as “Alcohol”, “Allergy,” “Fatigue,” “Coughing,” “Anxiety,” “Swallowing Difficulty,” “Wheezing,” “Chronic Disease,” “Peer Pressure,” and “Yellow Fingers.” Both of the models’ performances on the test data were acceptable with similar errors of 0.099 for ridge and 0.1 for lasso.

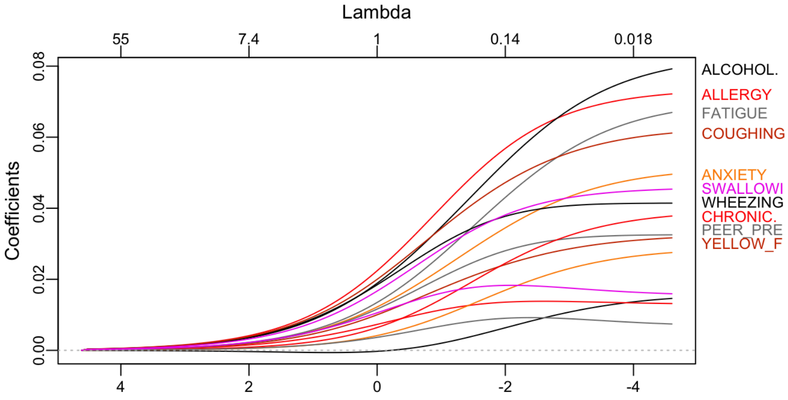


Figure 3:
Ridge Solution Path

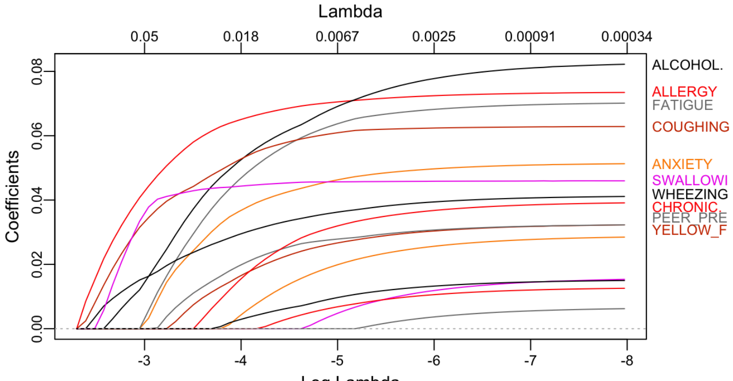


Figure 4:
Lasso Solution Path

Results

The first two methods are the logistic regression and KNN models. Based on the training and testing errors, after performing k value optimization, k = 7 would work best for the knn model. Logistic Regression and KNN models were fitted to the training dataset and predictions were made using the fitted models on both the training and testing datasets. The error rates for both Logistic Regression and KNN models were computed for both the training and testing datasets. As you can see in Figure 5, the logistic regression testing error rate was lower compared to the knn model indicating robustness and reliability, and is capable of making accurate predictions on

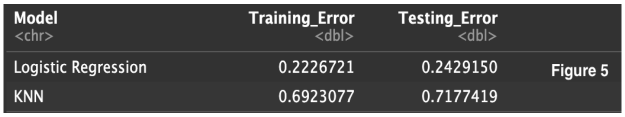

```
Model                 Training_Error   Testing_Error
<chr>                          <dbl>           <dbl>
Logistic Regression        0.2226721       0.2429150
KNN                        0.6923077       0.7177419
```

Figure 5

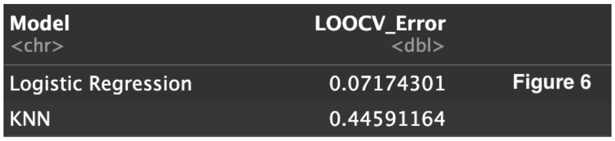

```
Model                      LOOCV_Error
<chr>                            <dbl>
Logistic Regression         0.07174301
KNN                         0.44591164
```

Figure 6

unseen data. Additionally, Leave-One-Out Cross-Validation was performed for both Logistic Regression and KNN models to estimate their generalization performance. Based on Figure 6, the LOOCV error for the logistic regression model was approximately 0.072. This means that, on average, when each data point is left out once and the model is trained on the remaining data points, it misclassifies about 7.2% of the data points that were left out. The LOOCV error for the KNN model was significantly higher at around approximately 0.446. This indicates that the KNN model, on average, misclassifies about 44.6% of the data points that were left out during cross-validation. The logistic regression model appears to have better predictive performance, as it had a lower testing error and a lower LOOCV error compared to the KNN model.

The third method is the decision tree. Using cross-validation, the best size of the tree turns out to be 12 (Figure 7), and the pruned tree with 12 nodes is shown in Figure 8. The decision tree displayed in Figure 7 utilized the important features in predicting lung cancer for splitting such as the variables “ALLERGY” , “SWALLOWING DIFFICULTY”, and “YELLOW FINGERS”. For example, if you were to follow the path of the allergy feature that is greater than or equal to 1.5, the decision tree suggests developing/having lung cancer. Fortunately, the decision tree model is easily interpreted and it presents a great prediction performance based on its low test error at 0.456.

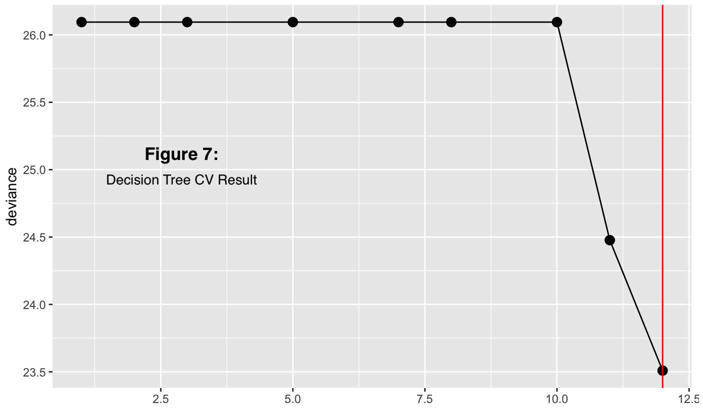
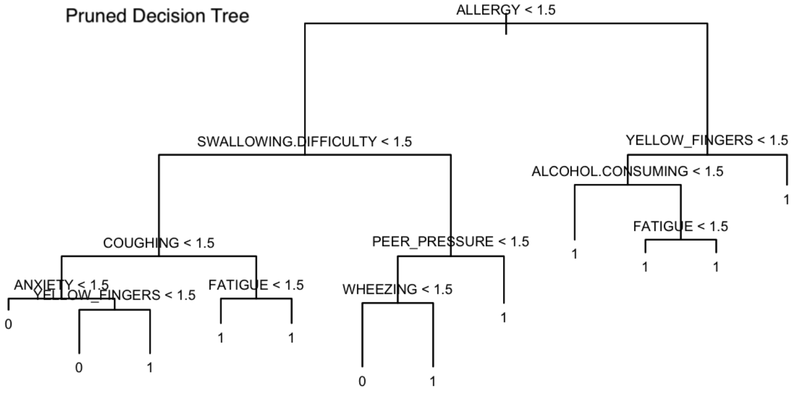


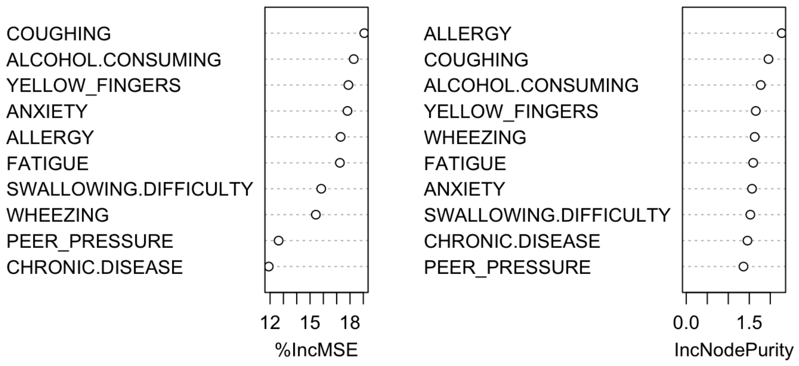
Figure 7:
Decision Tree CV Result

Figure 8:
Pruned Decision Tree



The fourth method is random forest, which is a tree-based model. When considering the feature importance indicators, “COUGHING” and “ALCOHOL.CONSUMING” are the most important features indicated by both Percent Increase in Mean Squared Error (%IncMSE) and Increase in Node Purity (IncNodePurity) as shown in Figure 9. Additionally, the test error of the random forest model is 0.1129, which indicates that it performs better compared to the logistic regression, KNN, and decision tree models.

Figure 9
Feature Importance Indicators



We cannot rely solely on the testing errors to make informed decisions about model selection for the best model fit. As a result, a ROC curve plot was produced to compare the AUC values for each model. Figure 10 presents the models and their AUC values. Amongst the models, Random Forest had the highest AUC (0.985) followed by Logistic Regression (0.966). Due to their high AUC values, confusion matrices were created to validate their performances in the ROC curve plot. Logistic regression had an accuracy rate of 91.9%, and random forest had an accuracy rate of 88.7%. These accuracy rates are evident that the two models are viable in accurately predicting lung cancer.

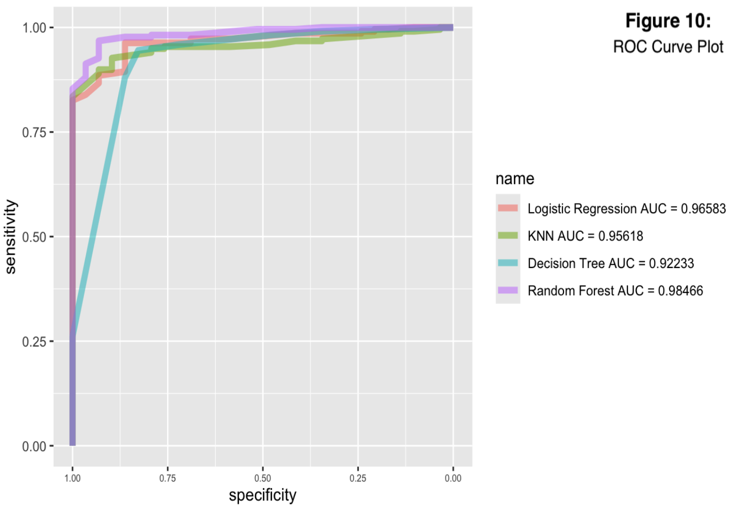


Figure 10:
ROC Curve Plot

Table of Model Performances

Model	Test Errors	AUC Values	Accuracy Rates	Sensitivity Values	Specificity Values
Logistic Regression	0.2429150	0.96583	0.9194	0.9808	0.6000
KNN	0.7177419	0.95618			
Decision Tree	0.4560601	0.92233			
Random Forest	0.1129032	0.98466	0.8871	0.9808	0.4000

Discussion

Given the clinical setting where prediction models could be used, it's important to have more than one prediction angle. Thus, out of all four models, both models, random forest and logistic regression, are optimal for predicting lung cancer. More specifically, the random forest model fits the training data best. However the logistic regression model had a higher accuracy and specificity, which is to correctly identify lung cancer detection.

In terms of the significant predictors of lung cancer diagnosis, the predictors showing importance in two out of the four models are “COUGHING”, “ALCOHOL.CONSUMING”, “ALLERGY”, and “YELLOW_FINGERS”. This is not a rigorously developed conclusion; rather, it is only a subjective conclusion suggesting that these could be the main areas of concern in future lung cancer diagnoses. Firstly, since the important predictors in this small dataset have been identified, it is worthwhile to think about creating a more diverse model with extra features like genetic markers, environmental factors, or additional clinical data and more observations in addition to the selected predictors from this dataset for further analysis. This may result in a more representative and diverse model, as well as test if the accuracy would hold up as the data grew in size. Secondly, both models reached a sensitivity rate of ~98%. That is to say, if applied in practical use, there's a risk for bias where a malignant tumor could be diagnosed as benign, which could be fatal. Further analysis and suggestions should include a more thorough hyperparameter optimization process by fine-tuning the random forest model and optimizing parameters such as the number of trees, max tree depth, and min samples per leaf to improve the model sensitivity. To sum up, the obtained models have pretty good performance in both the training and testing sets, high sensitivity and specificity are achieved.

References

1. Al Aswad, N. (2022, July 15). Lung Cancer DataSet. Kaggle. https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer
2. Centers for Disease Control and Prevention. (2023, June 8). *Lung cancer statistics*. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/lung/statistics/index.htm
3. *Lung cancer survival rates: 5-year survival rates for lung cancer*. 5-Year Survival Rates for Lung Cancer | American Cancer Society. (n.d.). https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html
4. Li, Y., Wu, X., Yang, P., Jiang, G., & Luo, Y. (2022). Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. *Genomics, proteomics & bioinformatics*, *20*(5), 850–866. https://doi.org/10.1016/j.gpb.2022.11.003