Business Problem

Lung cancer is a serious health issue that occurs when abnormal cells in the lungs grow out of control. Some workplace exposures can increase the risk of lung cancer, including exposure to hazardous chemicals, air pollution and pesticides:

- Chemicals: Exposure to harmful chemicals like arsenic, cadmium, chromium, diesel exhaust, nickel etc. can increase the risk of lung cancer. Smoking can also increase the risk of lung cancer. Some jobs like shipbuilders, refinery, insulation installers, tile workers also involve exposure to hazardous chemicals.
- Air pollution: Increasing levels of air pollution can also increase the risk of lung cancer.
- Pesticides: Occupational uses of pesticides can also increase the risk of lung cancer in some cases. Exposure of pesticides through skin contact and inhalation during application.

Background/History

Lung cancer has been a disease since at least the mid-19th century when doctors first described it. Before the 20th century, it was rare, but by the end of the century, it was the leading cause of cancer deaths in many developed countries. In the 21st century, it became the leading cause of cancer deaths worldwide, with 2.2 million cases and 1.8 million deaths in 2020. By 2012, it has surpassed breast cancer as the leading cause of deaths among women in developed countries. The rapid increase in the worldwide prevalence of lung cancer was attributed mostly to the increase use of cigarettes following world war I, though increases in environmental pollution were suspected to have been a contributing factor as well.

Lung cancer occurs primarily in people between the ages of 45 and 75 years. In countries with a prolonged history of tobacco smoking, between 80 and 90 percent of all cases are caused by smoking.

Tumors can begin anywhere in the lung, but symptoms do not usually appear until the disease has reached an advanced stage or spread to another part of the body. The most common symptoms include shortness of breath, a persistent cough or wheeze, chest pain, bloody sputum, unexplained weight loss, and lower respiratory infections. In cases where the cancer has spread beyond the lungs, visible lumps, jaundice or bone pain may occur.

Data Explanation (Data Prep/Data Dictionary/etc) – Done by shakti

Data Dictionary

Column Name	Data Type	Description
Patient_ID	Integer	A unique identifier for each patient.

Column Name	Data Type	Description
AGE	Integer	Age of the patient.
GENDER	String	Gender of the patient ('Male', 'Female').
SMOKING	String	Smoking status ('YES' if the patient smokes, 'NO' otherwise).
YELLOW_FINGERS	String	Indicates if the patient has yellow fingers ('YES'/'NO').
ANXIETY	String	Reports if the patient suffers from anxiety ('YES'/'NO').
PEER_PRESSURE	String	Indicates if peer pressure affects the patient's smoking habits ('YES'/'NO').
CHRONIC_DISEASE	String	Indicates the presence of any chronic disease ('YES'/'NO').
FATIGUE	String	Reports if the patient frequently feels fatigue ('YES'/'NO').
ALLERGY	String	Indicates if the patient has allergies ('YES'/'NO').
WHEEZING	String	Presence of wheezing symptoms ('YES'/'NO').
ALCOHOL_CONSUMING	String	Alcohol consumption status ('YES'/'NO').
COUGHING	String	Indicates if the patient has a cough ('YES'/'NO').
SHORTNESS_OF_BREATH	String	Patient experiences shortness of breath ('YES'/'NO').
SWALLOWING_DIFFICULTY	String	Difficulty in swallowing ('YES'/'NO').
CHEST_PAIN	String	Reports chest pain ('YES'/'NO').
LUNG_CANCER	String	Lung cancer diagnosis ('YES'/'NO').

Data Preparation

Data Cleaning and Transformation:

- Missing Values: Identified and handled missing data, either by imputation (replacing missing values with statistical estimates) or by removing records with incomplete information, depending on the amount and nature of the missing data.

- Categorical Encoding: Non-numeric categories such as 'YES/NO' responses, gender, or categorical risk factors were encoded into numeric formats suitable for analysis. This typically involves converting strings to binary (0/1) or multi-class integers using methods like `pd.get_dummies()`.

- Normalization/Standardization: If the dataset includes continuous variables like age or biomarker levels, these might have been normalized or standardized to ensure consistent scale and to improve the performance of machine learning models.

Feature Engineering:

- Derived new features potentially relevant to predicting lung cancer outcomes, such as age groups or risk factor scores, to enhance model insights and predictive accuracy.

1. Categorical Conversion: The 'LUNG_CANCER' and 'AGE_GROUP' columns in the dataset are converted to categorical data types. This transformation is crucial for enabling the LightGBM model to handle these features correctly, especially since 'LUNG_CANCER' is the target variable and 'AGE_GROUP' might be used as a basis for stratification.

2. Train-Test Split: The dataset is divided into training and testing sets, with 20% of the data reserved for testing. This split helps in validating the model's performance on unseen data, ensuring that the evaluation metrics reflect the model's ability to generalize.

Model Training

3. Group K-Fold Cross-Validation: The training data undergoes a Group K-Fold cross-validation process, specifically stratified by 'AGE_GROUP'. This method ensures that each fold is a good representative of the whole, especially considering the age distribution, which is critical for medical datasets like lung cancer where age can be a significant factor.

- GroupKFold: This cross-validation technique preserves the percentage of samples for each class and ensures that the same group is not represented in both testing and training sets. It is particularly useful when data within the same group has similar characteristics that could leak information about the target variable.

4. LightGBM Model Configuration: The model is configured with specific hyperparameters aimed at optimizing performance for binary classification ('binary' objective). Key parameters include:

- `learning_rate`, `max_depth`, `num_leaves`: Control the learning process and complexity of the model to prevent overfitting.

- `colsample_bytree`, `reg_alpha`, `reg_lambda`: Provide regularization, which is essential to avoid overfitting to the training data.

- `class_weight='balanced'`: Addresses class imbalance by adjusting weights inversely proportional to class frequencies.

5. Model Fitting and Evaluation: For each fold, the model is trained and evaluated:

- Early Stopping: Used to stop training if the validation score does not improve for 100 rounds, enhancing efficiency and preventing overfitting.

- Performance Metrics: The model's performance is evaluated using accuracy and F1-score, providing a balanced view of model performance, particularly in how well the model handles class imbalances.

- Confusion Matrix: Displays the model's performance in terms of true positives, false positives, true negatives, and false negatives, which is crucial for clinical settings where different types of errors have significantly different implications.

Post-Processing and Final Evaluation

6. Aggregation of Results: After training across folds, predictions (out-of-fold predictions) and true values are aggregated from all folds to provide a comprehensive view of the model's performance.

7. Final Metrics Calculation: Using the aggregated predictions and true values, final metrics such as overall accuracy and F1-score are calculated, summarizing the model's effectiveness across all cross-validation folds.

8. Visualization: The confusion matrix for each fold is plotted, providing a visual interpretation of where the model succeeds and where it makes errors.

Analysis

Overview of the Methodology

1. Data Setup:

- The dataset `df` is preprocessed by converting 'LUNG_CANCER' and 'AGE_GROUP' to categorical types to facilitate their proper handling by the LightGBM classifier.

- The data is split into training and testing sets, with 20% of the data reserved for testing to evaluate the model's generalization capability.

2. Cross-Validation Setup:

- Group K-Fold cross-validation (GroupKFold with 4 splits) is utilized, stratifying the folds based on 'AGE_GROUP' to ensure that each fold is representative of the overall age distribution. This strategy helps mitigate sampling bias and improves the model's robustness.

3. LightGBM Model Configuration:

- The LightGBM model is configured with binary classification settings and hyperparameters tuned for balance and efficiency, such as learning rate, max depth, number of leaves, regularization parameters, and the number of estimators. The class weight is set to 'balanced' to account for any class imbalances in the dataset.

Fold-wise Analysis

Training and Validation at Each Fold:

- Fold Initiation: Each fold starts with a setup log, indicating the training and validation sizes, ensuring transparency in data distribution across the folds.

- Model Training: The model is trained on the training subset of each fold. The validation subset is used to monitor the model's performance, specifically using accuracy as the metric. Early stopping is employed to halt training if no improvement is seen in 100 rounds, preventing overfitting and optimizing training time.

- Model Evaluation: At the end of each fold, the model's performance is evaluated on the validation set using accuracy and F1 score, providing a balanced view of model performance across precision and recall. These metrics are crucial in clinical settings where both the correct identification of positives (true positives) and the minimization of false alarms (false positives) are important.

Visualization and Interpretation:

- Confusion Matrix: Post-evaluation, a confusion matrix is displayed for each fold, visually summarizing the true positives, false positives, true negatives, and false negatives. This matrix is crucial for understanding the model's performance nuances, such as its ability to correctly identify lung cancer cases versus falsely identifying healthy cases as cancerous.

- Plot Display: The confusion matrix for each fold is plotted, providing immediate visual feedback on the classification accuracy across the binary outcomes.

Aggregated Results and Final Evaluation

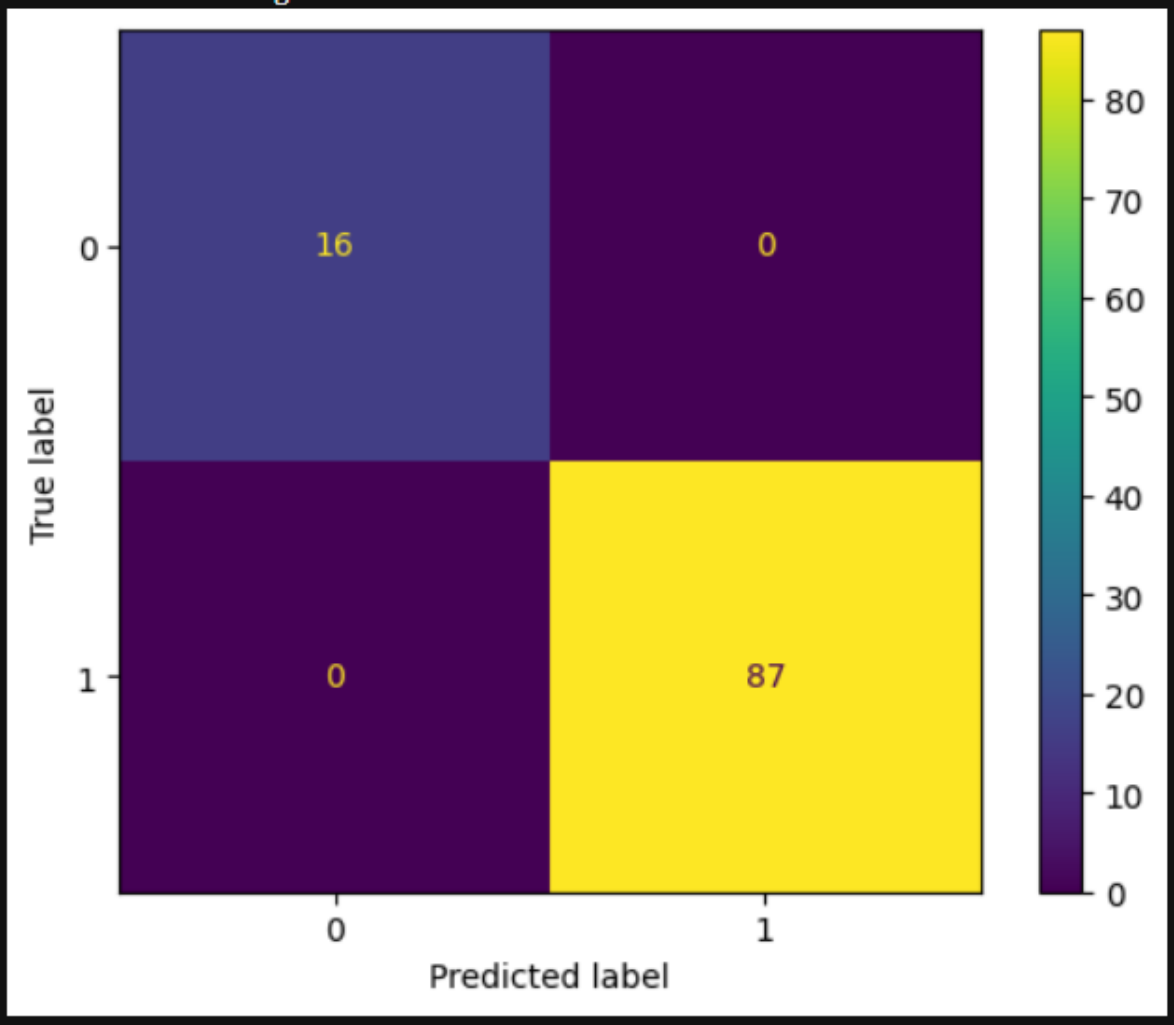
Aggregation of Predictions:

- After processing all folds, predictions and true values from each fold are aggregated to provide a comprehensive view of the model's performance across the entire cross-validation process.

Final Performance Metrics:

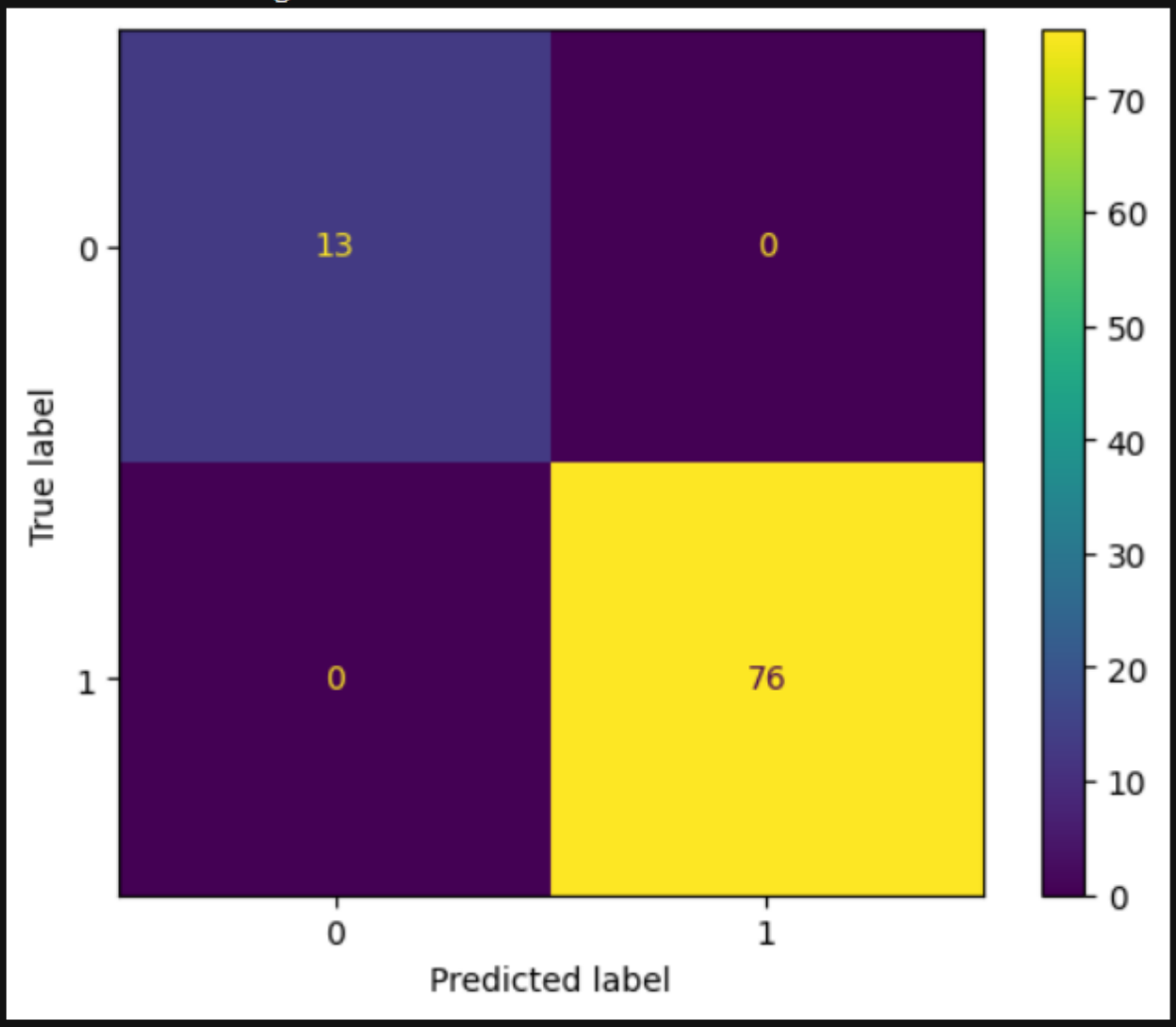
- Using the aggregated predictions and true values, final metrics such as overall accuracy and F1-score are computed. These metrics reflect the model's global performance, summarizing its effectiveness in predicting lung cancer across varied age groups and differing data subsets.

```
#########################
### Fold 1
### train size 144, valid size 103
#########################
Training until validation scores don't improve for 100 rounds
Did not meet early stopping. Best iteration is:
[1022]  valid_0's binary_logloss: 0.00318435
Accuracy for LightGBM =  1.0
F1 Score for LightGBM =  1.0
```

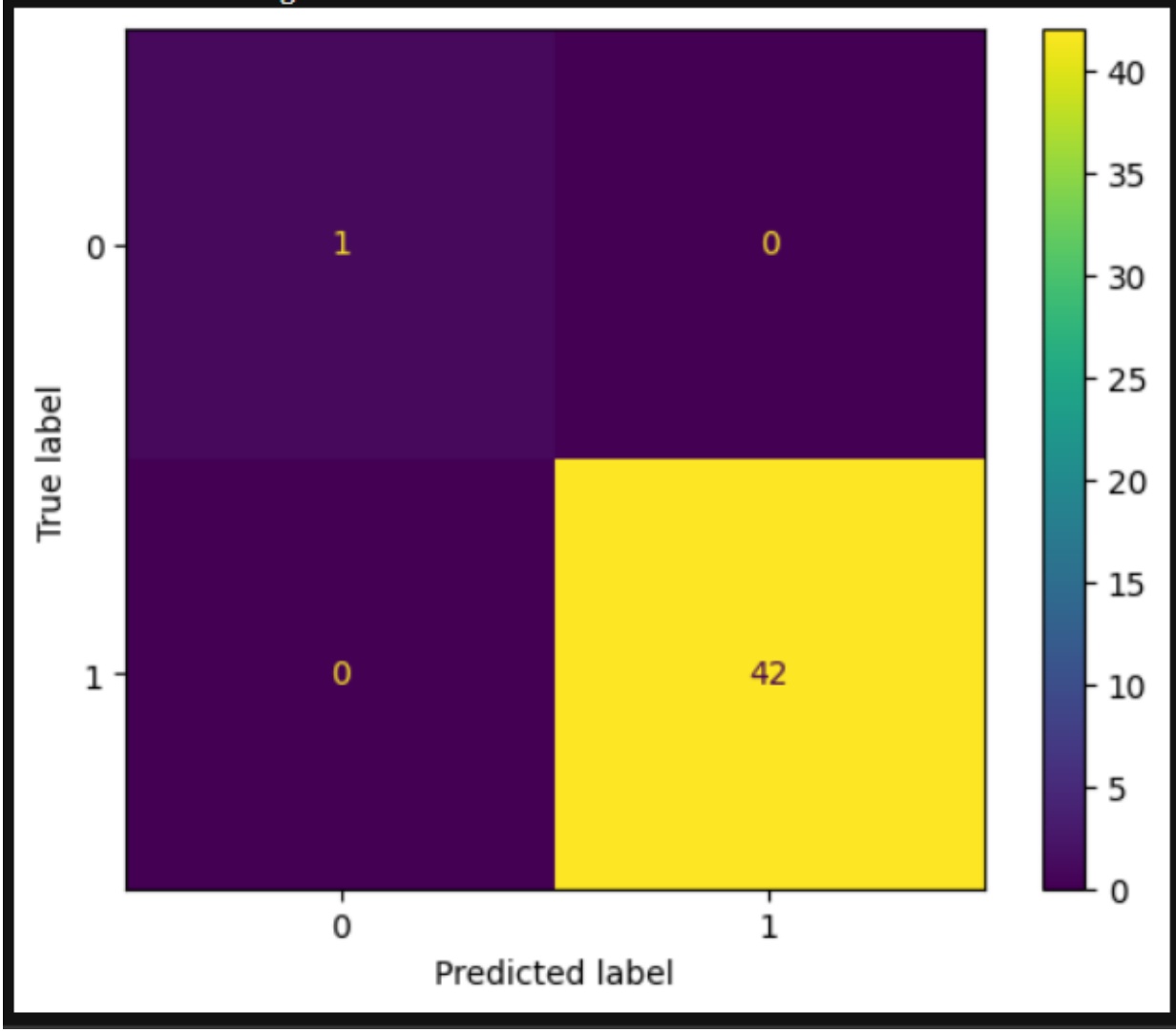



```
#############################
```

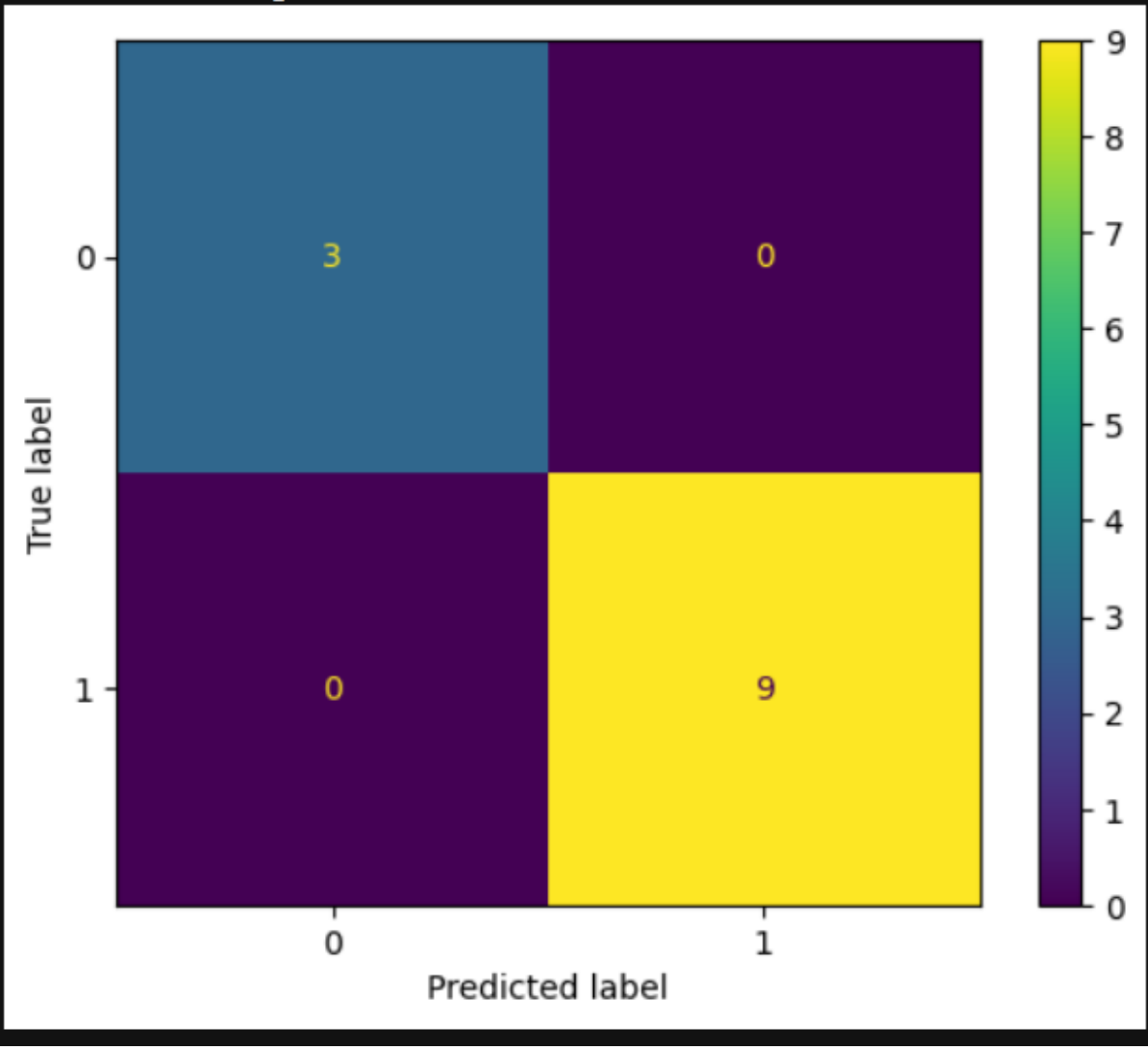
```
##########################
### Fold 2
### train size 158, valid size 89
##########################
Training until validation scores don't improve for 100 rounds
Did not meet early stopping. Best iteration is:
[1022]  valid_0's binary_logloss: 0.00305988
Accuracy for LightGBM =  1.0
F1 Score for LightGBM =  1.0
```

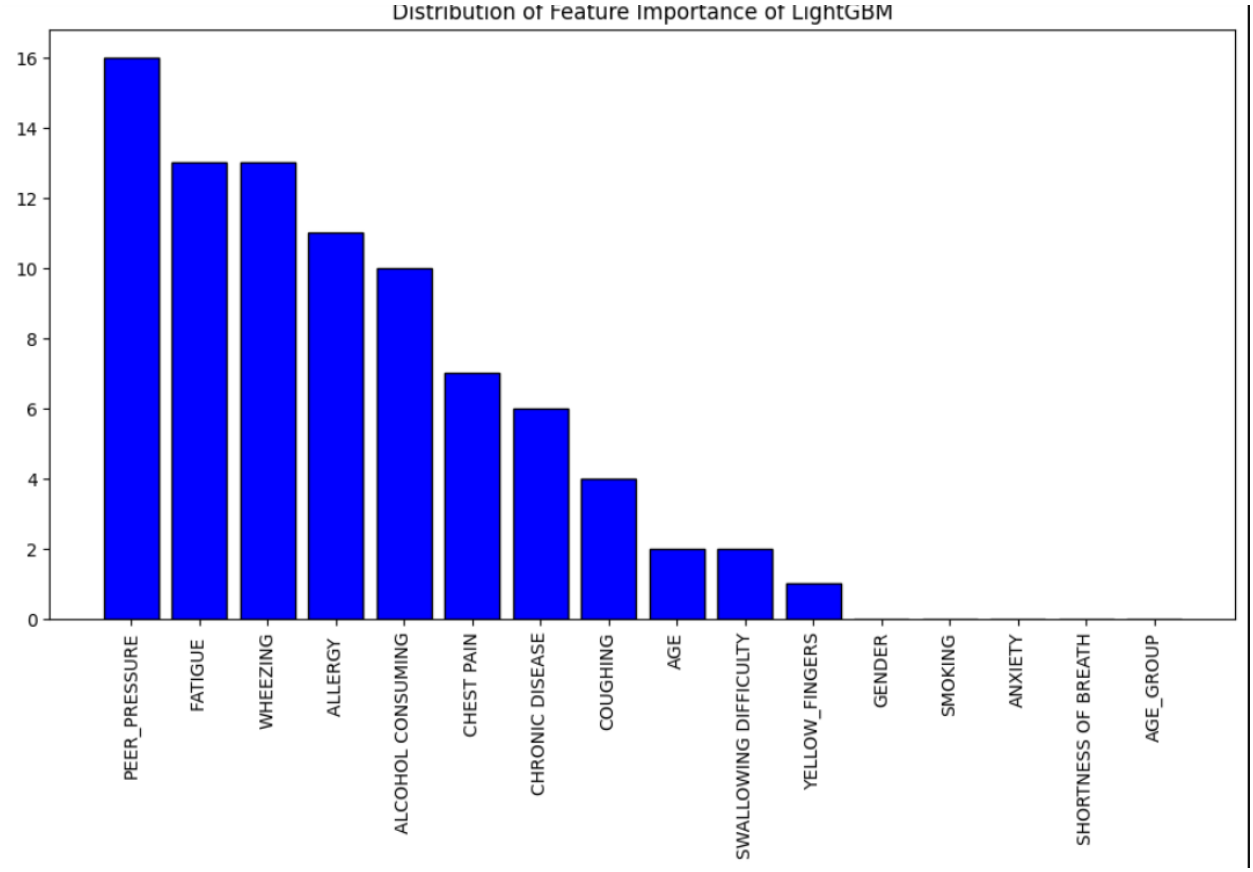


```
#############################
### Fold 3
### train size 204, valid size 43
#############################
Training until validation scores don't improve for 100 rounds
Did not meet early stopping. Best iteration is:
[1022]  valid_0's binary_logloss: 0.00184976
Accuracy for LightGBM =  1.0
F1 Score for LightGBM =  1.0
```



```
#########################
### Fold 4
### train size 235, valid size 12
#########################
Training until validation scores don't improve for 100 rounds
Did not meet early stopping. Best iteration is:
[1022]  valid_0's binary_logloss: 0.00155215
Accuracy for LightGBM =  1.0
F1 Score for LightGBM =  1.0
```





Interpretation of the Visualization

Key Insights:

- Feature Prioritization: The most important features (those with the highest bars) are likely candidates for closer examination in further analyses. These features are the most influential in determining whether a patient is predicted to have lung cancer based on the model's learning from the training data.

- Model Optimization and Simplification: Features with very low importance scores might be considered for removal in model simplification efforts, potentially speeding up the model and reducing overfitting without significantly compromising performance.

- Strategic Decisions: Understanding which features are most important can guide clinical focus areas, resource allocation, or further data collection strategies, particularly in medical or healthcare contexts where understanding risk factors or predictors of lung cancer is critical.

Utility of the Visualization:

- This visualization serves as a critical tool for stakeholders (e.g., data scientists, clinicians, healthcare policy makers) to understand the underlying drivers of the model's predictions. This can enhance trust and transparency in machine learning applications in sensitive fields like healthcare.

Conclusion

Lung cancer is the leading cause of death all over the world and the only chance of cure for patients affected from its kind of cancer is surgical resection.

This is mainly due to the fact that several factors are involved in lung cancer development and progression and to date the diagnostic methods available for early and efficient detection are not sufficient.

Although lung cancer research data has accumulated dramatically during the past several years, there is no database specifically focusing on lung cancer molecular biology available yet. Lung cancer may cause no symptoms until later stages, although some people may experience a chronic cough, changes to the voice and fatigue. People with high risk of developing lung cancer can consider regular screening. This can help detect the early signs and allow for treatment before the cancer spreads to the other part of the body.

Lung cancer has one of the lowest five-year survival rates of all cancers, with the national average in the United States being 15.6% - 26.6%. This is because lung cancer is often diagnosed at later stages, when it's less likely to be curable. Survival rates vary depending on the stage of cancer.

- Stage 1: 70-58% Survival rate
- Localized: 61.2% Survival rate
- Regional: 33.5% Survival rate
- Distant: 7% Survival rate

However, lung cancer has been improved significantly in recent years, with a better understanding of underlying genomics. Targeted therapies and immune therapies have improved outcomes for many patients.

Assumptions

Some People assume that lung cancer is incurable, especially if they are diagnosed early. However, there's a good chance of curing early stage of lung cancer with proper and timely treatment. Even for stage 4 lung cancer, there are treatments that can extend life and improve symptoms.

Smoking is one of the major causes of lung cancer. Cigarette smoking is full of cancer-causing substances called carcinogens. When you inhale cigarette smoke, the carcinogens cause changes in the lung tissue almost immediately. With each repetitive exposure, health cells that line your lungs become more damaged. Over time, the damage causes cells to change and eventually cancer may develop.

There are also few misconceptions about lung cancer including:

- Lung cancer is self-inflicted: Smoking is highly addictive and disproportionately affects vulnerable populations.
- There's no way to screen for lung cancer: Lung cancer can sometimes be found during a chest x-ray done for another condition.
- Lung cancer doesn't cause any symptoms until it is advanced.
- Some people assume that if they don't smoke, they don't have to worry about lung cancer: According to the American Cancer Society, about 20 percent of Americans died of lung cancer and they had never smoked or used any kind of tobacco.
- A lung cancer diagnosis is the end of life: Fatalism is a common thinking in lung cancer to the extent that many patients believe that even an early-stage diagnosis means the end of life.
- Lung cancer only happens to old people: It is a misconception that someone who is younger cannot get lung cancer. Lung cancer in younger people is rare but it happens.
- There's been no progress in lung cancer research: Early-stage lung cancer patients who undergo surgery now have effective platinum-based chemo combinations and metastatic patients have new tools like oral targeted therapies and immunotherapies. And many of these newer options have fewer side effects.

Another assumption that can lead to misdiagnosis is that patients with lung cancer have pneumonia. Doctors may assume that patients who have had frequent bouts of pneumonia in the past have it again, rather than checking for other problems.

Limitations

Lung cancer can cause several limitations, including:

- Shortness of breath: When lung cancer grows and block airways or causes fluid to build up around the lungs, it can make it difficult to breathe. This is called dyspnea. Lifestyle changes can help manage shortness of breath, such as quitting smoking, avoiding passive smoking, drinking plenty of water, resting practicing deep breathing and eating healthy foods. After lung cancer surgery, breathing may also be limited, but the remaining lung usually adapts over time with exercise and a healthy lifestyle.
- Muscle weakness: Small cell lung cancer can cause muscle weakness, particularly in the lower extremities, which can make everyday tasks even more difficult.
- Trouble Swallowing: Lung cancer can spread to the esophagus, making swallowing difficult.
- Heart conditions: Lung tumors can compress heart vessels, leading to serious heart conditions such as abnormal heart rhythms, fluid build up around the heart, or blockages that prevent blood from reaching the heart.
- Nerve pain: Tumors can cause nerve pain, as well as numbness, tingling and weakness.
- Fatigue and breathlessness: These side effects can cause physical and emotional changes and can also cause anxiety and depression.

- Screening Limitations: Low dose CT scan used for lung cancer screening can expose the lungs to radiation and may lead to overdiagnosis. It occurs when screening finds cancer that wouldn't cause illness or death, which can lead to unnecessary treatment and side effects.

- Post Surgery infections: Patients are at risk of infection after lung cancer surgery including wound, chest or urine infections. Antibiotics can help reduce the risk of infections but patients still experience some symptoms.

Challenges

Lung cancer can present several challenges like physical symptoms, emotional distress and socioeconomic instability for survivors.

- Physical Symptoms: Lung cancer can cause shortness of breath if it blocks major airways, or fluid to build up around the lungs and heart, making it harder to inhale. Other common symptoms like chest pain, coughing up blood, fatigue and weight loss.
- Emotional distress: Survivors may experience guilt, loneliness or worry that they burdened their caregivers and family. Counselling or a support group can help survivors express their feelings and overcome any guilt feeling.
- Socioeconomic instability: Survivors may face financial problems and sexual dysfunction. These issues can also be a challenge for the survivor in day-to-day life.

Improving survival rates is also another challenge for lung cancer patients, but there is opportunity for future research. This includes prevention, screening, surgery, radiotherapy and systematic therapies.

There are several barriers to lung cancer screening, including practical barriers like travel time and costs, and emotional barriers like fear, shame and stigma.

Lung cancer screening too faces several challenges, including selecting high risk individuals, training radiologists and optimizing screening intervals. Screening should also be accompanied by quality control to ensure radiation doses are optimized and that a number of positive screenings is expected. In the US, 14.9% of patients don't have a lung cancer screening center within 30 miles, which is especially prevalent in rural areas. Expanding telehealth coverage could help address this challenge. A significant challenge for the United States lung cancer screening program is poor uptake by low income but high-risk candidates. To optimize results from potential curative radical radiotherapy and surgery, accurate staging of patients is vital; modern staging can improve patient selection for radical treatment.

Risk factors: Lung cancer risk factors include environmental and occupational exposures, family history, genetic risks and exposure to outdoor air pollution.

Improving survival in lung cancer patients remains a challenge dependent on prevention, screening, optimal surgery, modern radiotherapy and improved systematic therapies targeted through understanding the molecular biology of these heterogeneous tumors. Despite clear progress to date, there is much need for improvement, offering ample opportunity for future research.

Future Uses/Additional Applications

Early detection and treatment of early-stage disease is the most effective way to save lives from lung cancer. Immunotherapy and targeted therapies represent some of the biggest advances in cancer treatment and have had a major impact on the approach for the treatment of lung cancer. While 20 years ago all lung cancers were treated the same, now specialists understand that lung cancer really means a multitude of diseases and targeted therapies can be applied to effectively treat tumors with certain genetic mutations.

Many NCI funded researchers at the NIH campus and across the United States and the world are seeking ways to address lung cancer more effectively. Some research is basic, exploring questions as diverse as the biological underpinnings of cancer and the social factors that affect cancer risk. And some are clinical, seeking to translate basic information into improved patient outcomes.

Newer therapies are available for patients with advanced lung cancer. These primarily include immunotherapies and targeted therapies, which continue to show benefits as research evolves.

ALK (**Anaplastic lymphoma kinase**) inhibitors: It targets cancer causing rearrangements in a protein called ALK. These drugs continue to be refined for the 5% of patients who have ALK gene alteration.

EGFR (epidermal growth factor receptor) inhibitors: It blocks the activity of a protein called epidermal growth factor receptor. Altered forms of EGFR are found at high levels in some lung cancers, causing them to grow rapidly.

ROS1 inhibitor: The ROS1 protein is involved in cell signaling and cell growth. A small percentage of people with lung cancer have rearranged forms of the ROS1 gene.

BRAF inhibitors: The B-Raf protein is involved in sending signals in cells and cell growth. Certain changes in the B-Raf gene can increase the growth and spread of lung cancer cells.

Clinical Trials: Trails are available for both non-small cell lung cancer treatment and small cell lung cancer treatment.

Recommendations

Lung cancer patients can follow certain tips to stay healthy during and after treatment.

- Continue to avoid tobacco: Quitting smoking can have immediate benefits.
- Eat well: Good nutrition is extremely important.
- Stay active: Physical activity can help with fatigue, weight and mood.
- Get Support: Address your mental health and emotional needs and have someone to take care of the patient.
- Manager side effects: Patient undergoing chemotherapy, targeted therapy or radiation therapy might experience nausea, vomiting, diarrhea and constipation. Ginger ale drink, peppermint tea, eating bland and simple food or eating smaller meals can help the patient to handle the side or aftereffects of the therapies.

The U.S. Preventive Services Task Force (USPSTF) recommends annual lung cancer screening with low dose computed tomography in people who meet all of these criteria: Are ages 50 to 80 years. Have a 20 pack-year smoking history. Currently smoke cigarettes or quit within the past 15 years.

Implementation Plan

A lung cancer screening program with low dose CT screening is a complex endeavor with the purpose of identifying persons without symptoms with lung cancer in an early stage allowing curative treatment, avoiding causing harm to the persons that do not have the disease. To achieve this during large scale implementation requires that the screening program is performed according to a systematic, structured, standardized and validated protocol, and that the quality of the performance is monitored continuously.

Aside from costs, healthcare resource capacity and access to screening programs need to be considered. Implementing a lung cancer screening program requires both CT scanner capacity, as well as the training of sufficient radiologists to evaluate the screens. An additional aspect to consider is the increased demand in surgical capacity due to screening. While late-stage lung cancers are generally treated through radiotherapy and chemotherapy, early-stage lung cancers are often treated through surgical resection. Given that lung cancer screening will lead to an increased detection of early-stage cancers, an increased demand for surgical capacity should be expected. Indeed, modelling studies for the U.S. have suggested that an adherence rate of 50% would require a 37% increase in surgical capacity. Therefore, health-care resource capacity should be taken into consideration for implementing lung cancer screening. Gradual implementation strategies with modelling informing resource allocation should be considered.

Ethical Assessment

The risk of lung cancer increases with age and cumulative exposure to tobacco, smoke, and decreases with time since quitting smoking. The best evidence for the benefit of screening comes from the NLST, which enrolled adults 55 to 74 years of age who had at least a 30 pack-year smoking history and were current smokers or had quit within the past 15 years. As with all screening trials, the NLST tested a specific intervention over a finite period. Because initial eligibility extended through 74 years of age and participants received three annual screening CT scans, the oldest participants in the trial were 77 years of age.

The USPSTF used modeling studies to predict the benefits and harms of screening programs that use different screening intervals, age ranges, smoking histories, and times since quitting. A program that annually screens adults 55 to 80 years of age who have a 30 pack-year smoking history and currently smoke or quit within the past 15 years is projected to have a reasonable balance of benefits and harms. The model assumes that people who achieve 15 years of smoking cessation during the screening program discontinue screening. This model predicts the outcomes of continuing the screening program used in the NLST through 80 years of age.

Questions asked by audience:

1- How much experience do specialists have in treating lung cancer?
2- How quickly does a patient have to decide on treatment?
3- Will treatment affect daily activities?

4- How long will the treatment last? What will treatment be like?
5- How does the patient know if the treatment is working?
6- What are the symptoms and the side effects?
7- Are there any limitations for the patient?
8- How often does a lung cancer patient need to have follow ups and imaging tests?
9- What are the options if the cancer comes back?
10- What should be the goal of the treatment?

References:

- https://www.medicalnewstoday.com/articles/323701#causes
- https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620
- https://www.masseycancercenter.org/news/the-lung-cancer-care-of-the-future
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7947387/
- https://www.aafp.org/pubs/afp/issues/2014/0715/od1.html