Department Of Computer Science and Engineering

Lung Cancer Prediction

Course Title: Artificial Intelligence and Expert Systems Lab

Course Code: CSE 404

Date of Submission:26/04/2025

Submitted To:

Noor Mairukh Khan Arnob

Lecturer,

Department of CSE, UAP

Submitted By:

Name: Swapna Khanam

Reg no: 21201082

Sec: B2

1.Introduction

Lung cancer is a critical public health issue and a major cause of cancer-related mortality globally. According to the World Health Organization, it accounts for approximately 1.8 million deaths annually. One of the main challenges with lung cancer is its late detection due to the subtlety of early symptoms. Early diagnosis can drastically increase survival chances.

This project aims to address this issue by developing a machine learning model capable of predicting the likelihood of a lung cancer diagnosis based on various symptoms and lifestyle factors gathered through survey data. Such a model can assist in creating cost-effective, accessible early screening tools.

2. Problem Statement

The goal is to develop a classification model that can:

- Accurately predict whether a person is at risk of lung cancer.
- Utilize survey-based inputs such as age, gender, smoking habits, and symptoms.
- Assist in early diagnosis through an automated, data-driven approach.

3. Dataset Overview

The dataset used for this project is titled **survey lung cancer.csv**, which contains survey responses from individuals. Each record includes demographic information, lifestyle choices, common lung cancer symptoms, and the final diagnosis.

Key Features in the Dataset:

Feature	Description
GENDER	Gender of the individual (Male/Female)
AGE	Age in years
SMOKING	Whether the individual smokes (0/1)
YELLOW_FINGERS	Presence of yellow-stained fingers (0/1)
ANXIETY, FATIGUE, ALLERGY, WHEEZING	Symptoms (0/1)
ALCOHOL CONSUMING	Alcohol usage (0/1)
COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, CHEST PAIN	Key symptoms (0/1)
LUNG_CANCER	Diagnosis label (Yes/No, target variable)

4. Exploratory Data Analysis (EDA)

EDA was conducted to better understand the structure and patterns within the dataset.

Key Observations:

- **No missing values** were present, which simplified preprocessing.
- **Class distribution**: There were more positive lung cancer diagnoses than expected, hinting at some bias in the survey design.
- **Symptom prevalence**: Features like smoking, yellow fingers, and wheezing were common among those diagnosed.

Visualizations:

- **Countplot of Lung Cancer Cases**: Helped visualize the class imbalance.

- **Correlation Heatmap**: Used to find relationships between symptoms and cancer likelihood. Strong positive correlations were found between smoking, coughing, and cancer diagnosis.

5. Data Preprocessing

Before model training, several preprocessing steps were applied:

- **Label Encoding**: Converted categorical values like GENDER and LUNG_CANCER into numerical format.
- **Feature Scaling**: Applied StandardScaler to normalize the feature values and ensure better convergence in logistic regression.
- **Train-Test Split**: Data was split into training (80%) and testing (20%) sets using train_test_split from scikit-learn.

This process ensured that the model receives clean and standardized input for accurate learning.

6. Model Selection and Training

For this project, the **Logistic Regression** algorithm was chosen because:

- It is simple and interpretable.
- It works well for binary classification problems.
- It provides probability estimates, which can help in risk assessment.

Training Process:

- The model was trained on the scaled training dataset.
- Predictions were generated on the test dataset using the predict() method.

7. Model Evaluation

To assess the performance of the logistic regression model, several evaluation metrics were used:

Metric	**Description**
Accuracy	Overall percentage of correct predictions
Precision	Ratio of correctly predicted positive observations to total predicted positives
Recall	Ratio of correctly predicted positive observations to all actual positives
F1 Score	Harmonic mean of precision and recall

A **confusion matrix** was also plotted to visualize the number of true positives, true negatives, false positives, and false negatives.

Performance Summary:

- The model demonstrated **good accuracy** and balance between precision and recall.
- Most cancer cases were correctly identified, showing strong potential for real-world application.

8. Results

The logistic regression model performed well on the test data, accurately predicting lung cancer based on simple survey inputs. Key findings include:

- **Smoking**, **coughing**, and **chest pain** were the strongest indicators of lung cancer.
- The model was effective even without using advanced clinical data like imaging or lab results.

9. Future Work

To enhance this system, future improvements can include:

- **Advanced Models**: Testing more sophisticated algorithms such as Random Forest, XGBoost, or Neural Networks.
- **Feature Engineering**: Incorporating additional data like patient history, imaging, or genetic data.
- **Dealing with Imbalanced Data**: Using techniques like SMOTE to improve model robustness.
- **Deployment**: Creating a simple web or mobile app for public use or integration into hospital systems.

10. Conclusion

This project successfully demonstrated the application of machine learning in the medical domain, particularly for **early detection of lung cancer** using basic, non-invasive inputs.

Key Takeaways:

- Logistic regression offers a simple, interpretable, and efficient solution for binary classification.
- With minimal features, a reliable predictive system was developed.
- Such tools can help in pre-screening, saving time and resources in healthcare.