Academic Year	Module	Assessment Number
2A	Concepts and Technologies of AI	3

Lung Cancer Classification Analysis Report

Student Id : 2438466

Student Name : Anup Maharjan

Section : L5CG22

Module Leader : Mr. Siman Giri

Tutor : Mr. Ronit Shrestha

Submitted on : 11-02-2025

Table of Contents

Table of Figures

Abstract

This report will attempt to predict a categorical variable from a set of classifications (whether a person had lung cancer).

The dataset used for this analysis is therefore the Lung Cancer Survey dataset that includes indicators of health and statical data for 309 individuals. The process is: Exploratory Data Analysis (EDA), building model with Logistic Regression & Decision Trees, Feature selection & Hyper parameter optimization.

Accuracy, precision, recall and F1 score were used to evaluate the performance of the models.
Positive cases have an accuracy of 92.86% and F1 score of 0.96 using Logistic Regression model.

However, the classification models performed very well, especially the Logistic Regression beating the Decision Trees. It turns out to be the most predictive features are chronic disease, fatigue, coughing, allergy and swallowing difficulty.

1 Introduction

1.1 Problem Statement

This project aims to predict a particular health indicator, and demographic features are present if lung cancer is also present or absent in the patient. The goal of this classification task is to provide assistance in early detection and risk assessment of lung cancer by using characteristics of patient and observable symptoms.

1.2 Dataset

In this analysis, we were using the Lung Cancer Survey dataset which consists of responses of 309 people. This includes demographic information as well as a number of health indicators and symptoms. This is a dataset that fits into United Nations Sustainable Development Goal 3 (Good Health and Well-being) contributing to early warning and risk reduction of non-communicable diseases.

1.3 Objective

The objective of this analysis is to develop a predictive model which predicts lung cancer outcomes through the dataset based on provided features.

2 Methodology

2.1 Data Preprocessing

The data was cleaned prior to building the model 33 duplicate entries were removed and also categorical variables were converted to Boolean format. The dataset was free of any missing values. Also, the data were prepared to be analyzed by applying transformations, such as feature standardization.

2.2 Exploratory Data Analysis (EDA)

Plots such as Bar Plot, Histogram, Box Plot, Scatter Plot and Heat Map was performed for EDA.

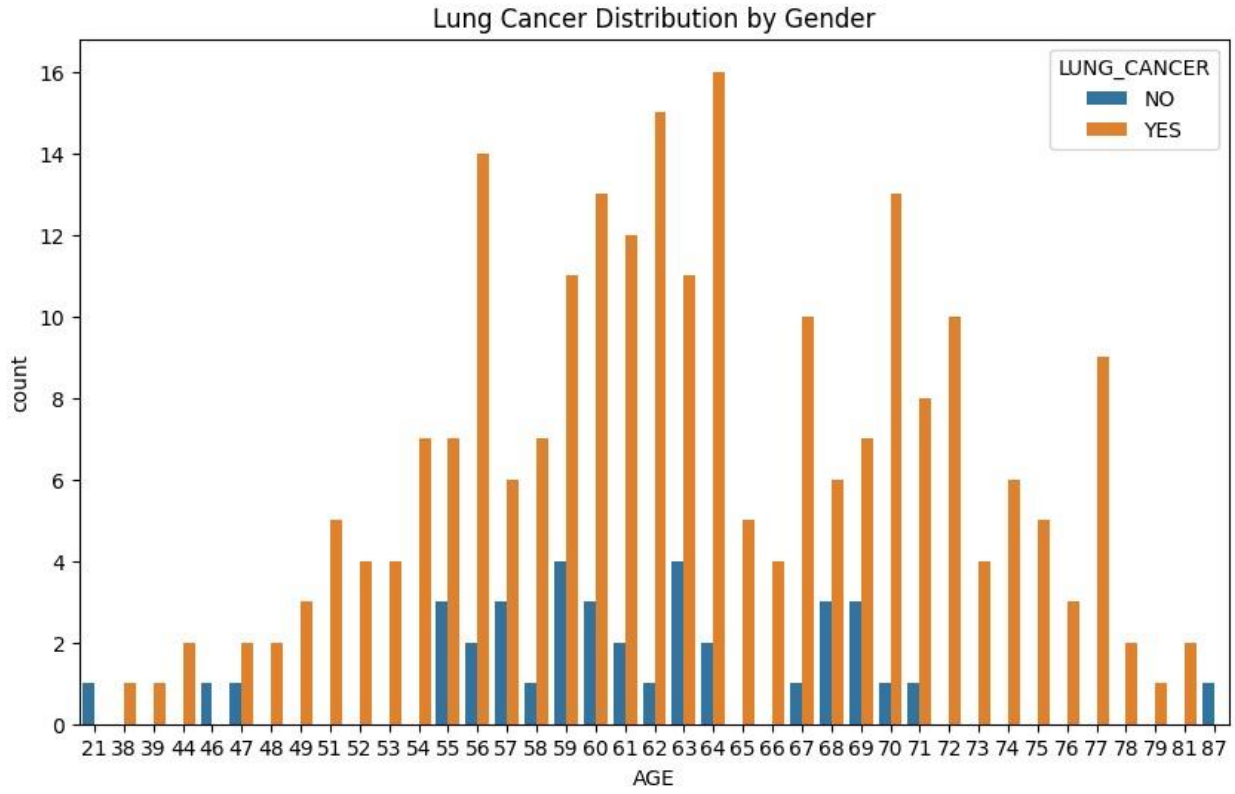

Figure 1: Bar Plot

The bar chart above shows the distribution of lung cancer cases across different ages, ranging from 21 to 87 years old. The highest number of cases (around 17.5) occurs in the mid-60s age range with a concentration of cases between ages 50-75. The number of cases tends to be lower at both younger ages (below 40) and older ages (above 80)

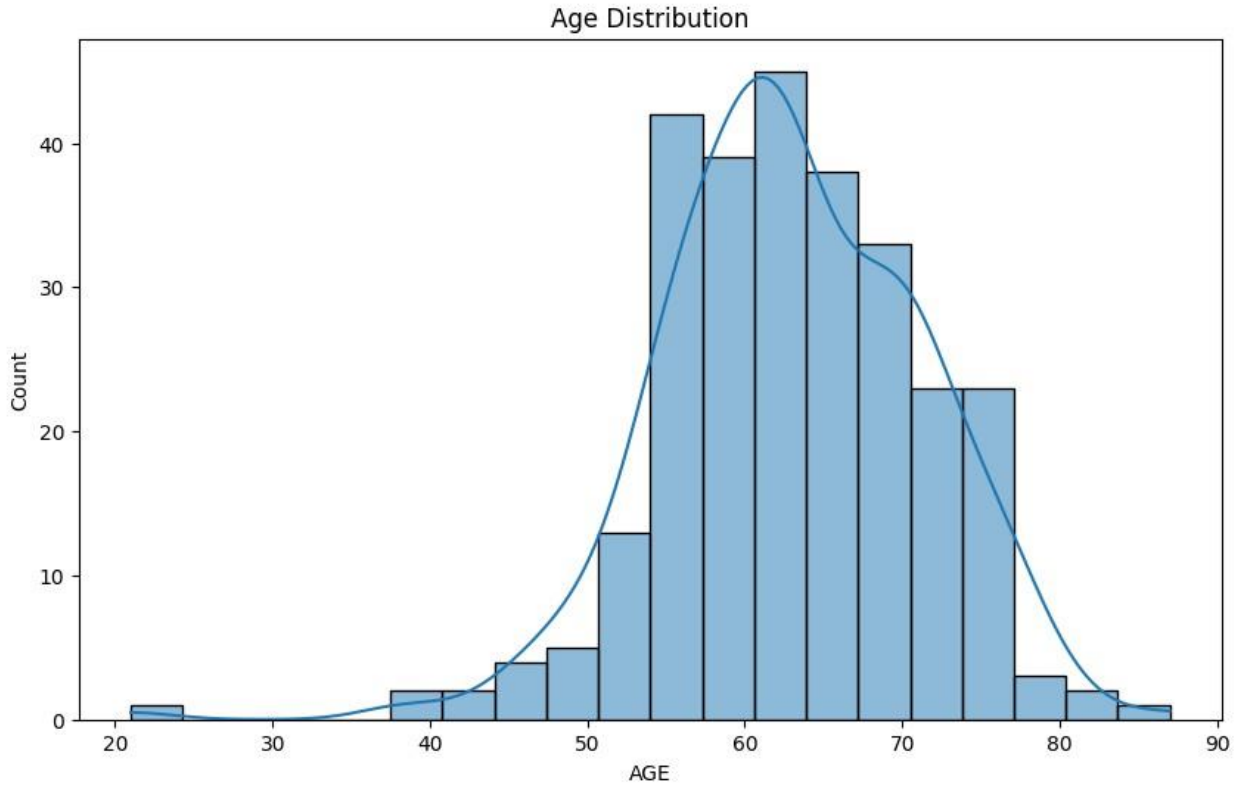

Figure 2: Histogram Plot

The above histogram shows the age distribution range from approximately 20 to 90 years with a clear peak in the distribution between ages 60-65, with the highest count reaching about 50-55 individuals. There is Left-Skeweness / Negatively Skeweness, the tail is longer on the left side (towards younger ages), indicating that fewer young people data exist.The peak (mode) is around 60 years, and the distribution declines more gradually towards older ages. Most cases are concentrated between ages 55-75

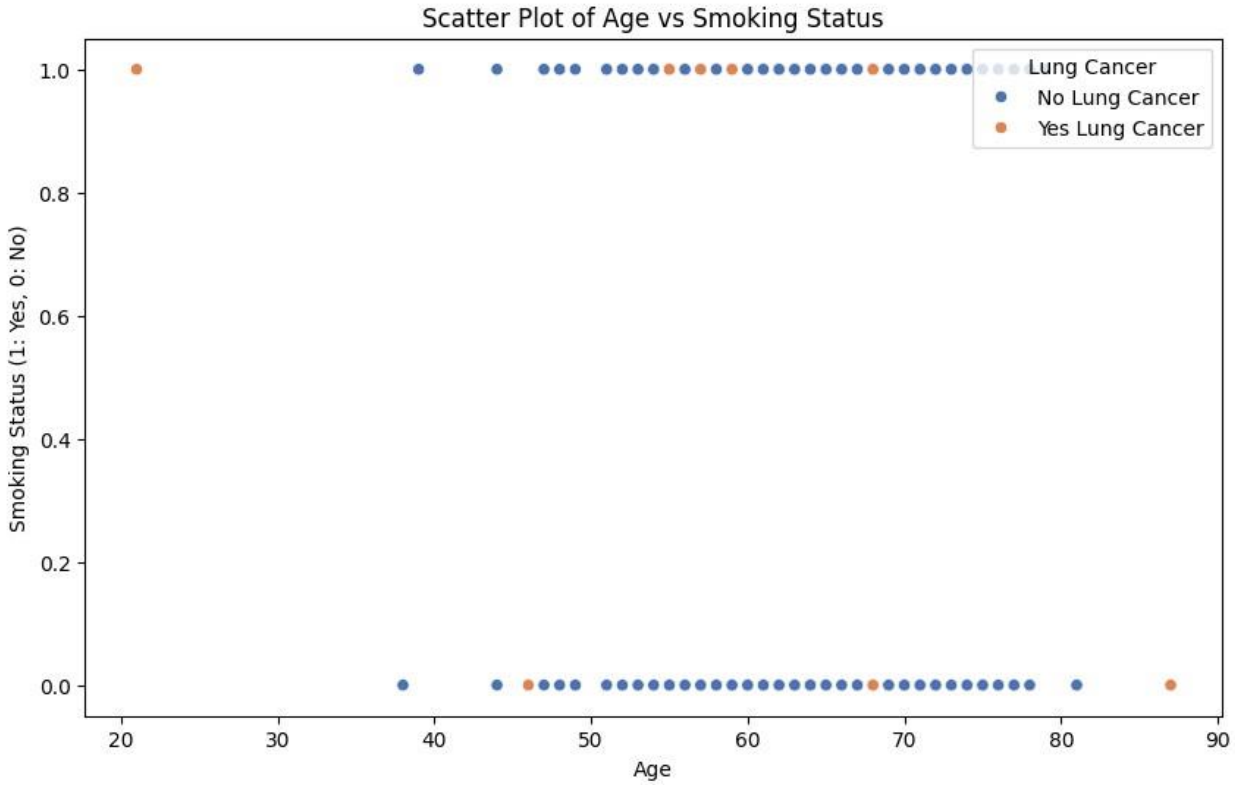

Figure 3: Scatter Plot

The above scatter plot shows the relationship between age and smoking status, and the points are colored according to lung cancer status. The spread points over ages from about 20 to 90 years, and, clearly, separate between smokers (1.0 in y-axis) and non-smokers (0.0 in y-axis). People without lung cancer is represented by blue dots, people with lung cancer is represented by orange dot. There appears to be a higher concentration of points between ages 50-80 Both smokers and non-smokers can be found with and without lung cancer, though there seems to be more orange dots (lung cancer cases) among the smokers (1.0 level)

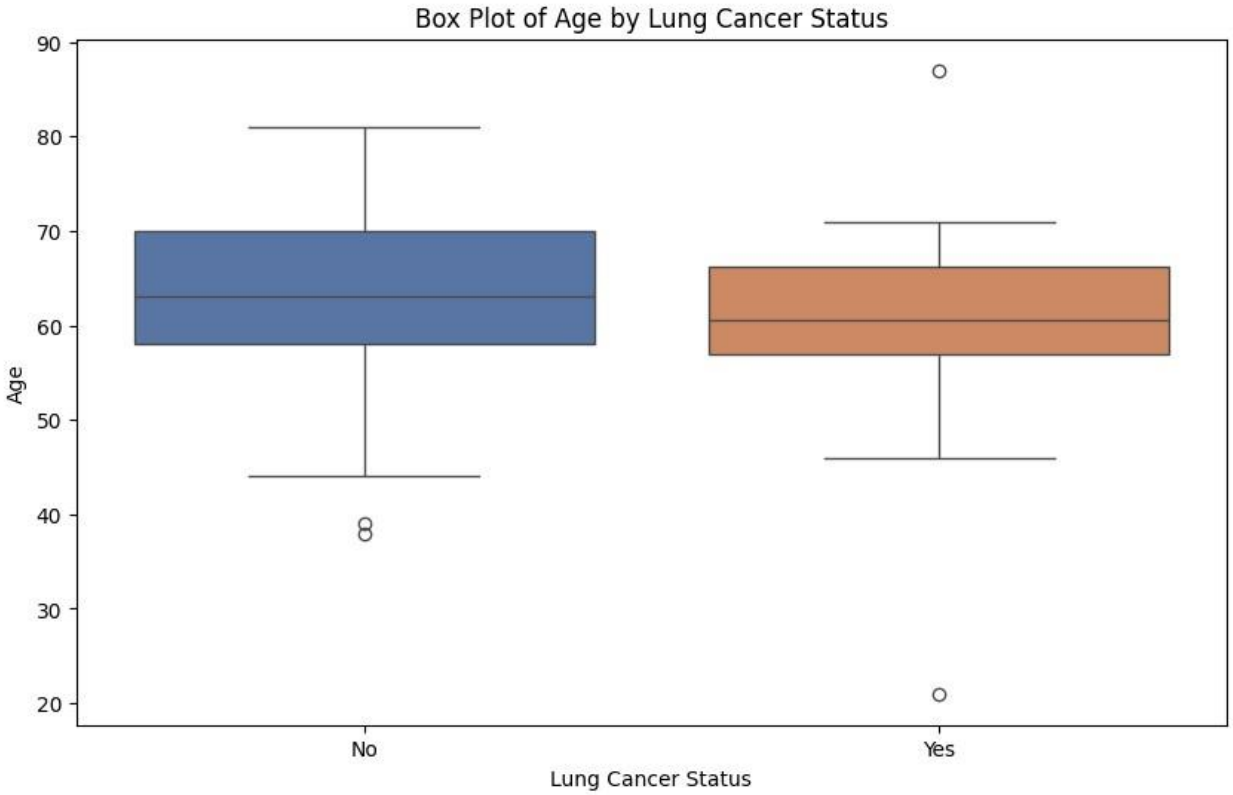

Figure 4: Box Plot

The median age (shown by the horizontal line in each box) is almost similar for both groups, around 60-65 years. Both groups show almost similar interquartile ranges, suggesting similar age spread in the middle 50% of cases. There are some outliers in both groups: for "No" lung cancer: outliers around age 40 and for "Yes" lung cancer: outliers at both very young (around 20) and very old (around 85) ages. The distributions appear symmetrical for both groups. Overall, this visualization suggests that the age distributions are quite similar between people with and without lung cancer in this dataset.

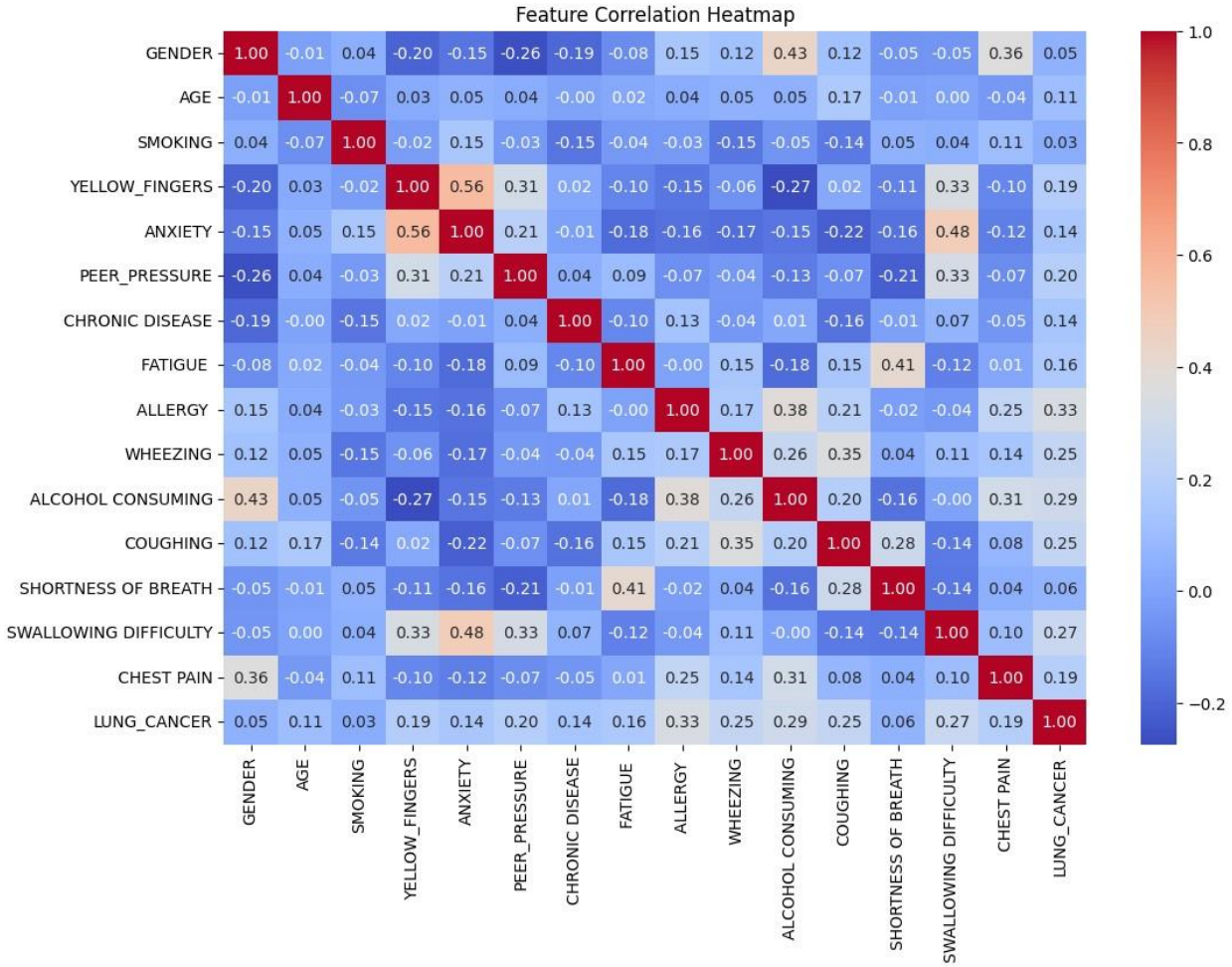

Figure 5: Heat Map

The above heatmap shows the correalation and relationships between various features related to lung cancer.

Strong positive correlations:

- Yellow fingers and anxiety (0.57)
- Swallowing difficulty and anxiety (0.49)
- Shortness of breath and fatigue (0.44)
- Wheezing and coughing (0.37)
- Allergy and alcohol consuming (0.34)

Moderate correlations with lung cancer:

- Allergy (0.33)
- Alcohol consuming (0.29)
- Coughing (0.25)
- Wheezing (0.25)

- Swallowing difficulty (0.26)

Weak or negligible correlations:

- Gender and lung cancer (0.09)
- Age and lung cancer (0.06)
- Smoking and lung cancer (surprisingly low at 0.06)

Negative correlations:

- Yellow fingers and alcohol consuming (-0.29)
- Peer pressure and shortness of breath (-0.22)
- Anxiety and coughing (-0.23)

Key insights of the EDA are:

- Age distribution is concentrated between 55-75 years
- Imbalanced dataset with 238 positive and 38 negative cases
- Strong correlations between certain symptoms (yellow fingers-anxiety: 0.57, swallowing difficulty-anxiety: 0.49)
- Moderate correlations between lung cancer and health indicators (allergy: 0.33, alcohol consuming: 0.29)

2.3 Model Building

For this task, two classification models considered were Logistic Regression and Decision Trees. Training and test split was employed on the data using 80% and 20% of the cases, for which preprocessed data was used to build custom versions of both algorithms and train them.

2.4 Model Evaluation

The two model's performance was evaluated based on some of the factors like: accuracy of prediction, how precise the model can be, Recall (the correct prediction with correct identifications), F1-Score. These choices of metrics have also been chosen because it is common to use them as a metric to evaluate classification models when having imbalanced classes.

2.5 Hyper-parameter Optimization

For better performance of the model both Logistic Regression and Decision Trees were run on GridSearchCV and RandomizedSearchCV. The optimal parameters were:

- Logistic Regression: C=1, max_iter=100
- Decision Tree: max_depth=15, min_samples_split=10, min_samples_leaf=4

2.6 Feature Selection

In order to discover key characteristics for lung cancer prediction Recursive Feature Elimination (RFE) performed automatic feature selection. The selected features were:

- For Logistic Regression
    1. Chronic Disease
    2. Fatigue
    3. Allergy
    4. Coughing
    5. Swallowing Difficulty

- For Decision Tree:
    1. Chronic Disease
    2. Age
    3. Allergy
    4. Alcohol Consuming
    5. Swallowing Difficulty

3 Conclusion

3.1 Key Findings

The model evaluated dataset based on Accuracy, precision, recall and F1 score. With a 0.96 F1-score and 92.86% accuracy for positive cases, the results were excellent.

3.2 Final Model

Finally, a Logistic Regression with chosen features-based model was the most effective to predict the presence of lung cancer. The test set accuracy corresponded to 92.86%.

3.3 Challenges

Several challenges occurred while working in the project such as class imbalance in the dataset and surprising low correlations for some expected predictors such as smoking.

3.4 Future Work

Further work for improving the model could be by collecting more balanced data, introducing more pertinent features, and looking into ensemble methods.

4 Discussion

4.1 Model Performance

Different comprehensive metrics were used to evaluate the performance of the model. Finally, the results clearly show that the model made an incredible job in the test data specially with predicting positive cases.

4.2 Impact of Hyperparameter Tuning and Feature Selection

To improve the model's performance hyperparameter tuning and feature selection played an important role. When we apply these techniques in the model, it increased the accuracy from 78.57 up to 92.86%.

4.3 Interpretation of Results

The selected features and model respectively worked as expected. They identified the key insights, which are that some symptoms and health conditions are strong predictors of lung cancer risk.

4.4 Limitations

There were some limitations such as small dataset and class imbalance, although there were limitations the model showed a good performance

4.5 Suggestions for Future Research

The use of more advanced feature engineering techniques, collecting other relevant health indicators and further investigation of application of ensemble methods are future research.