

**UNIVERSITY OF CALCUTTA**



**BASANTI DEVI COLLEGE**

**STUDY ON LUNG CANCER DATASET FOR**  
**BOTH MALE AND FEMALE PATIENTS**

PROJECT WORK DSE-B2 BY **SREYA BHATTACHARYA**

C.U. REGISTRATION NO. **041-1211-0278-20**

ROLL NUMBER **203041-11-0050**

Under guidance of **Prof. Dithi Bhattacharya**

Submitted for Bachelors of Science Course (Honours) of Department of  
STATISTICS

Under CBCS system for year 2023

**DECLARATION OF ORIGINAL WORK:-**

This declaration is made on:

Student's Declaration:

I, Sreya Bhattacharya, Roll Number - 203041-11-0050

hereby declare that the work submitted for the module DSE-B2 is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgment is made explicitly, nor has any part been authored by another person.

Date submitted --

Received for examination by: Calcutta University

Date:

• Examiner Signature :

(DITHI BHATTYACHARIYA)

• Examinee signature:

Student Name :

(SREYA BHATTACHARYA)

## **CONTENT:**

	Page
Chapter-1	
Executive Summary	4
Chapter-2	
Introduction	5-7
Chapter-3	
Dataset	08-14
Source of Data	14
Data Description	15
Chapter-4	
Methodology	16-19
Chapter-5	
Analysis	20-28
Chapter-6	
Conclusion	29
Chapter-7	
References	30
Chapter-8	
Acknowledgement	31
Chapter-9	
Annexure	32-36

# Chapter – 1

## **EXECUTIVE SUMMERY:**

Worldwide lung cancer is the second most commonly diagnosed cancer with an estimated 2.21 million of new cases every year (according to WHO). Cancer is a disease in which cells in the body grow out of control. Lung Cancer begins in the lungs and may spread to lymph nodes or other organs in the body such as brain, liver or any other organ also may spread through the lungs.

As per Lung Cancer Research Foundation, An estimated 238,340 People will be diagnosed with lung cancer in 2023 in the US. 1 in 16 People will be diagnosed with lung cancer in their lifetime (1 in 16 men, and 1 in 17 women.).

I have collected a dataset of prediction of Lung Cancer based on some symptoms like Gender, Age, Smoking, Yellow fingers, Anxiety, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing, Difficulty, Chest pain.

We will check which are the most sufficient factors among these symptoms and will work with those. Here we will fit a Multiple Logistic Regression model because Lung Cancer is the dependent variable and I'm studying the effect that the independent variables (i.e. the symptoms) have on the probability of obtaining a particular value of the dependent variable.

We will check how accurately our data fits in our Model.

## Chapter - 2

### INTRODUCTION

Lung Carcinoma is a malignant lung tumour characterized by uncontrolled cell growth in tissues of the lung. If left untreated, this growth can spread beyond the lung by the process of metastasis into nearby tissue on other parts of the body.

#### **Incidence:**

According to National Cancer Institute (A part of the U.S. Department of Health and Human Services), the rate of new cases of lung and bronchus cancer was 50.0 per 100,000 men and women per year. The death rate was 35.0 per 100,000 men and women per year. These rates are age-adjusted and based on 2016–2020 cases and deaths.



As per National Cancer Institute this graph is portraying how rate of new cases of lung cancer per 100,000 persons is varying in the years from 1992 to 2020.

#### **Symptoms:**

Different people have different symptoms for lung cancer. Lung cancer symptoms may include –

- Coughing that gets worse or doesn't go away.
- Chest pain.
- Shortness of breath.
- Wheezing.
- Coughing up blood.
- Feeling very tired all the time.
- Weight loss with no known cause.

*Most people with lung cancer don't have symptoms until the cancer is advanced.*

## Types of LUNG CANCER:

The two general types of lung cancer include:

- Small cell lung cancer
- Non-small cell lung cancer

## Risk Factors:

A number of factors may increase risk of lung cancer. Some risk factors can be controlled, for instance, by quitting smoking. And other factors can't be controlled, such as family history.

Risk factors for lung cancer include:

- Smoking.
- Previous radiation therapy
- Exposure to radon gas
- Exposure to asbestos and other carcinogens.
- Family history of lung cancer

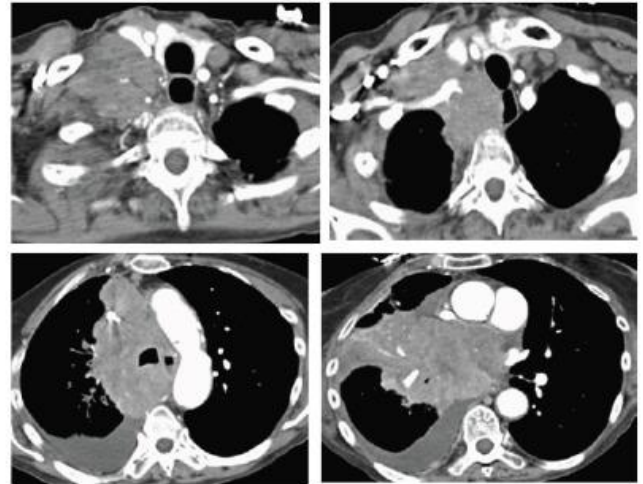
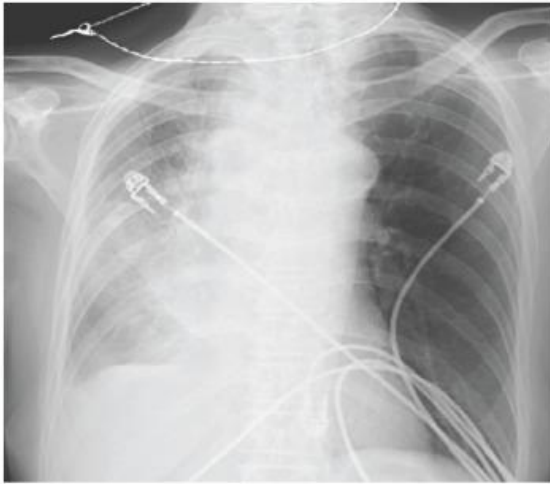
## Prevention:

There's no sure way to prevent lung cancer, but one can reduce the risk by following this steps.

- **Stop smoking:** Quitting reduces risk of lung cancer, even if one has smoked for years.
- **Test your home for radon:** If one live in an area where radon is known to be a problem, then he should test his home for radon.
- **Avoid carcinogens at work:** One should take precautions to protect himself from exposure to toxic chemicals at work. For instance, always wear face mask for protection.
- **Eat a diet full of fruits and vegetables:** One should choose a healthy diet with a variety of fruits and vegetables. Food sources of vitamins and nutrients are best.
- **Exercise most days of the week:** If a person don't exercise regularly, he should start out slowly. And he should try to exercise most days of the week.

In this project I want to observe the effect of Smoking, Yellow fingers, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Shortness of Breath, Swallowing Difficulty, Chest pain on Lung Cancer. So I'll use Multiple Logistic Regression so that I can foresee whether the patient has Lung Cancer or not based on the symptoms.

I have collected this X-ray and CT scan plate from the official website of National Institute of Health.



The left one portrays X-ray and the right one portrays CT scan of a lung cancer positive patient.

## Chapter : 3

### Data Set:

I've divided the whole dataset in Male and Female.

- **Table for Female :**

GENDER	AGE	SMOKING	ANXIETY	.....	CHEST_PAIN	LUNG_CANCER
Female	21	1	0		0	0
Female	38	0	0		1	1
Female	44	1	1		0	1
Female	47	1	0		0	1
Female	48	0	1		0	1
Female	49	0	0		0	1
Female	49	0	1		0	1
Female	51	1	1		0	1
Female	51	1	1		0	1
Female	51	1	1		0	1
Female	53	1	1		1	1
Female	53	0	1		0	1
Female	54	1	0		1	1
Female	54	1	1		0	1
Female	54	1	1		0	1
Female	54	1	1		1	1
Female	54	1	1		1	1
Female	55	1	1		0	0
Female	55	1	0		1	1
Female	55	0	0		1	1
Female	55	1	1		0	1
Female	55	1	1		0	1
Female	56	0	0		0	0
Female	56	0	0		1	1
Female	56	0	0		1	1
Female	56	0	1		1	1
Female	56	1	1		1	1
Female	56	0	0		0	1
Female	56	0	1		1	1
Female	56	1	1		1	1
Female	56	0	0		1	1
Female	57	1	0		0	0
Female	57	1	0		1	0
Female	57	0	1		0	0
Female	57	0	0		0	1
Female	57	0	1		1	1
Female	58	0	0		0	0
Female	58	1	1		0	1



Female	59	0	0	1	0
Female	59	1	0	0	0
Female	59	1	1	0	1
Female	59	0	1	0	1
Female	59	1	1	1	1
Female	59	0	1	0	1
Female	59	0	1	0	1
Female	59	0	1	0	1
Female	60	1	0	0	0
Female	60	0	1	0	0
Female	60	1	0	0	1
Female	60	0	0	0	1
Female	60	1	1	1	1
Female	60	0	0	1	1
Female	60	1	1	0	1
Female	61	0	0	0	0
Female	61	1	1	0	1
Female	61	1	1	1	1
Female	61	1	1	1	1
Female	61	1	1	0	1
Female	61	0	1	0	1
Female	61	1	0	0	1
Female	61	0	0	1	1
Female	61	1	1	1	1
Female	61	1	1	1	1
Female	62	0	0	0	0
Female	62	1	1	1	1
Female	62	1	1	0	1
Female	62	0	0	1	1
Female	62	1	0	1	1
Female	62	1	1	0	1
Female	63	0	0	0	0
Female	63	0	0	0	0
Female	63	0	0	1	1
Female	63	1	1	1	1
Female	63	0	0	1	1
Female	64	1	0	0	0
Female	64	0	1	0	1
Female	64	1	0	1	1
Female	64	0	0	0	1
Female	64	0	1	0	1
Female	64	1	1	1	1
Female	64	1	0	0	1
Female	64	1	1	1	1
Female	65	1	1	0	1
Female	65	0	1	0	1

Female	66	1	1	0	1
Female	66	1	1	0	1
Female	67	0	0	0	0
Female	67	1	1	1	1
Female	67	1	1	0	1
Female	67	1	1	0	1
Female	67	1	1	0	1
Female	68	1	1	0	0
Female	68	1	1	0	1
Female	68	0	1	0	1
Female	68	0	0	1	1
Female	69	0	1	0	1
Female	69	0	0	0	1
Female	70	1	0	0	0
Female	70	1	0	0	1
Female	70	0	1	1	1
Female	70	0	0	1	1
Female	70	0	0	0	1
Female	70	0	0	0	1
Female	70	1	0	0	1
Female	71	1	0	0	0
Female	71	1	1	1	1
Female	71	0	0	1	1
Female	71	1	1	1	1
Female	71	1	1	1	1
Female	71	0	0	0	1
Female	72	1	1	1	1
Female	72	0	0	0	1
Female	72	0	0	0	1
Female	72	1	1	0	1
Female	72	0	1	0	1
Female	73	1	1	1	1
Female	73	1	0	1	1
Female	73	1	1	0	1
Female	74	0	1	0	1
Female	74	1	1	0	1
Female	75	0	0	0	1
Female	75	0	0	0	1
Female	75	0	1	0	1
Female	76	1	1	1	1
Female	76	0	1	1	1
Female	77	0	1	0	1
Female	77	0	1	1	1
Female	77	0	0	1	1
Female	77	0	0	0	1
Female	77	1	1	0	1

Female	78	1	1	0	1
Female	81	0	0	0	1
Female	81	0	1	1	1
Female	87	0	0	0	0

• **Table for Male:**

GENDER	AGE	SMOKING	ANXIETY .....	CHEST_PAIN	LUNG_CANCER
Male	63	1	1	1	0
Male	69	1	0	1	0
Male	69	0	0	1	0
Male	55	0	0	0	0
Male	56	1	1	0	0
Male	59	0	1	1	0
Male	60	0	1	0	0
Male	69	1	0	1	0
Male	47	1	0	1	0
Male	68	0	1	0	0
Male	63	1	0	0	0
Male	61	0	1	0	0
Male	59	0	0	1	0
Male	68	1	1	0	0
Male	64	1	0	1	0
Male	55	1	0	1	0
Male	46	0	1	1	0
Male	69	0	1	1	1
Male	74	1	0	1	1
Male	52	1	0	1	1
Male	53	1	1	1	1
Male	72	0	0	1	1
Male	58	1	0	1	1
Male	75	1	0	1	1
Male	57	1	1	1	1
Male	60	1	0	1	1
Male	72	1	1	1	1
Male	65	0	1	1	1
Male	62	1	1	1	1
Male	60	0	0	1	1
Male	56	1	0	1	1
Male	60	1	0	1	1
Male	68	1	1	1	1
Male	63	0	0	0	1

Male	52	1	0	1	1
Male	72	1	1	1	1
Male	62	1	0	1	1
Male	63	1	1	0	1
Male	49	1	0	1	1
Male	52	0	1	0	1
Male	73	0	0	1	1
Male	47	0	0	1	1
Male	69	1	1	1	1
Male	70	0	0	1	1
Male	70	0	0	0	1
Male	74	0	0	1	1
Male	66	1	0	0	1
Male	68	1	1	1	1
Male	67	0	0	1	1
Male	61	1	0	1	1
Male	58	1	0	0	1
Male	56	1	0	1	1
Male	67	1	1	1	1
Male	56	0	0	1	1
Male	60	0	1	0	1
Male	66	0	1	1	1
Male	62	0	1	1	1
Male	52	1	0	1	1
Male	48	1	0	1	1
Male	60	0	1	1	1
Male	59	1	0	0	1
Male	64	0	1	0	1
Male	64	1	0	1	1
Male	62	1	1	1	1
Male	53	0	0	1	1
Male	58	0	1	1	1
Male	61	1	1	1	1
Male	60	1	1	1	1
Male	57	1	0	1	1
Male	77	1	1	0	1
Male	64	0	0	1	1
Male	57	1	1	0	1
Male	70	1	0	1	1
Male	51	1	0	1	1
Male	58	1	1	1	1
Male	76	1	0	1	1
Male	71	1	1	1	1
Male	69	0	1	0	1
Male	67	0	0	1	1
Male	63	0	0	0	1

Male	62	1	1	1	1
Male	65	1	1	0	1
Male	64	0	1	1	1
Male	51	0	0	1	1
Male	70	1	0	1	1
Male	58	1	0	1	1
Male	67	1	1	1	1
Male	62	0	0	0	1
Male	75	1	1	1	1
Male	62	1	1	0	1
Male	68	1	0	1	1
Male	63	0	1	0	1
Male	62	0	0	1	1
Male	44	0	0	1	1
Male	56	1	1	1	1
Male	54	0	0	1	1
Male	56	0	0	0	1
Male	72	0	0	1	1
Male	64	1	0	0	1
Male	63	1	1	1	1
Male	71	0	0	1	1
Male	72	1	0	1	1
Male	77	1	0	0	1
Male	55	1	0	0	1
Male	65	1	1	0	1
Male	63	0	0	1	1
Male	69	0	1	0	1
Male	64	0	1	1	1
Male	59	0	1	1	1
Male	74	1	0	1	1
Male	79	1	0	1	1
Male	62	0	1	1	1
Male	67	0	1	0	1
Male	55	1	0	1	1
Male	54	1	0	0	1
Male	64	1	1	1	1
Male	70	0	0	1	1
Male	59	1	0	1	1
Male	71	0	1	0	1
Male	57	0	0	1	1
Male	78	0	0	1	1
Male	64	1	1	1	1
Male	62	0	0	1	1
Male	77	0	0	0	1
Male	63	0	0	1	1
Male	59	1	0	1	1

Male	77	1	1	1	1
Male	61	0	1	0	1
Male	67	0	0	0	1
Male	69	1	1	0	1
Male	59	0	1	1	1
Male	74	0	0	1	1
Male	64	1	0	1	1
Male	70	1	1	1	1
Male	58	0	0	0	1
Male	39	1	0	1	1
Male	70	1	0	1	1
Male	60	1	1	1	1
Male	55	0	0	1	1
Male	63	1	0	1	1
Male	56	1	0	1	1
Male	60	0	1	1	1

### **SOURCE OF DATA:**

- Kaggle Lung Cancer Dataset.  
Link: <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>
- The website of online lung cancer prediction system.

## **DATA DESCRIPTION:**

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status.

This data consists,

Total no. of attributes: 16

No. of instances: 276 (for male 142 and for female 134)

### Attribute information:-

1. Gender : Male, Female
2. Age : Age of the patient
3. Smoking : YES=1 , NO=0
4. Yellow fingers : YES=1 , NO=0
5. Anxiety : YES=1 , NO=0
6. Peer pressure : YES=1 , NO=0
7. Chronic Disease : YES=1 , NO=0
8. Fatigue : YES=1 , NO=0
9. Allergy : YES=1 , NO=0
10. Wheezing : YES=1 , NO=0
11. Alcohol Consuming : YES=1 , NO=0
12. Coughing : YES=1 , NO=0
13. Shortness of Breath : YES=1 , NO=0
14. Swallowing Difficulty : YES=1 , NO=0
15. Chest pain : YES=1 , NO=0
16. Lung Cancer : YES=1 , NO=0

Here I divided my dataset into two parts i.e. - Male (142 observations) and Female (134 observations). I'll check the accuracy separately.

## Chapter : 4

### METHODOLOGY:

#### Multiple Logistic Regression Model

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

Binary Logistic Regression is used when the response is binary (i.e., it has two possible outcomes). The cracking example given above would utilize binary logistic regression. Other examples of binary responses could include passing or failing a test, responding yes or no on a survey, and having high or low blood pressure.

**Basic assumptions that must be met for logistic regression** include

Independence of errors, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers.

The multiple **binary logistic regression model** is the following:

$$\begin{aligned}\Pi(x) &= \exp (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k) / [1 + \exp (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)] \\ &= \exp (X\beta) / [1 + \exp (X\beta)] \\ &= 1 / (1 + \exp (-X\beta))\end{aligned}$$

The logistic regression model can be written as follows,

$$\text{Log} [\Pi(x) / (1 - \Pi(x))] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Where,

$X_j$ : The  $j$  th predictor variable.

$\beta_j$ : The coefficient estimate for the  $j$  th predictor.

Here  $\pi$  denotes a probability, that an observation is in a specified category of the binary.

- $Y$  variable, generally called the "success probability."  
(Notice that the model describes the *probability of an event* happening as a function of  $X$  variables. For instance, it might provide estimates of the probability that an older person has heart disease).
- With the logistic model, estimates of  $\pi$  from equations like the one above will always be between 0 and 1. The reasons are:

The numerator  $\exp (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$  must be positive, because it is a power of a positive value ( $e$ ). The denominator of the model is  $(1 + \text{numerator})$ , so the answer will always be less than 1.



With one  $X$  variable, the theoretical model for  $\pi$  has an elongated "S" shape (or sigmoidal shape) with asymptotes at 0 and 1, although in sample estimates we may not see this "S" shape if the range of the  $X$  variable is limited.

Now we discuss about the concepts of different measurement and how they are used to analyse the data.

### **p-value :**

A p-value is a statistical measurement used to validate a hypothesis against observed data. The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical **Hypothesis Test**, assuming that the **Null Hypothesis** is correct. The lower the p-value is, greater the statistical significance of the observed difference. Generally, p value less than 0.05 is considered statistically significant.

### **McFadden's $R^2$ :**

McFadden's  $R^2$  is used in model checking which measures the efficiency of the model. McFadden's  $R^2$  is defined as,

$$1 - LL_{\text{mod}} / LL_0$$

Where,  $LL_{\text{mod}}$  is the log likelihood value for the fitted model and  $LL_0$  is the log likelihood for the null model which includes only an intercept as predictor.

McFadden's  $R^2$  ranges between 0 to 1. The value near to 0 indicates the model's low predictive power and value over 0.40 indicates that the model fit the data very well.

### **Variable Importance :**

Variable importance or VI represents the statistical significance of each variable in the data with respect to its effect on the generated model. VI actually refers to how much a variable contributes in a given model to make accurate predictions. The more a model relies on a variable to make accurate prediction, the more the variable is important for that particular model.

Variable importance is calculated by the sum of the decrease in error when split by a variable. Then, the relative importance is the variable importance divided by the highest variable importance value so that values are bounded between 0 and 1.

### **VIF :**

In a regression model the covariates should be independent. If the covariates become dependent, the problem is called multicollinearity problem. VIF or Variance Inflation Factor measures the multicollinearity of a model and it is defined as,

$$VIF_j = 1 / (1 - R_j^2)$$

Where,  $R_j^2$  is the coefficient of determination of regression of  $X_j$  (a particular covariate) on  $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_p$  (on other covariates).

As a thumb rule, VIF value less than 5 indicates presence of less multicollinearity in the model. But VIF value greater than 5 indicates severe multicollinearity.

### Confusion Matrix :

Confusion matrix describes the performance of a classification model on a set of test data. It shows our predictions compared to the actual class. This matrix contains 4 elements --- TP, FP, TN and FN.

True Positive (TP): Observation is predicted positive and is actually positive.

False Positive (FP): Observation is predicted positive but is actually negative.

True Negative (TN): Observation is predicted negative and is actually negative.

False Negative (FN): Observation is predicted negative but is actually positive.

The confusion matrix can be written as follows,

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

### Sensitivity :

Sensitivity or the TP rate (TPR) is the fraction of positive values out of the total actual positive instances, i.e., the proportion of actual positive cases that are correctly identified.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Sensitivity lies between 0 to 1. Greater value indicates that the model can predict the data ver well.

### Specificity :

Specificity gives the fraction of negative values out of the total actual negative instances. In other words, it is the proportion of actual negative cases that are correctly identified. The FP rate is given by (1 – specificity).

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Specificity also lies between 0 to 1. The more the specificity is the more model can predict the data well.

### Accuracy :

Accuracy gives the proportion of the total number of predictions that are correct. In other words, it is the fraction of actual cases (both positive and negative) out of all cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

The total misclassification error rate can be defined as  $(1 - \text{accuracy})$ . The less the total misclassification error rate is or the more the accuracy is, the more the particular model predicts the data very well.

### **AUC-ROC Curve :**

The receiver operating characteristics (ROC) curve is the plot between sensitivity and the FP rate for various threshold values. The area under curve (AUC) is the area under this ROC curve. It is used to measure the quality of a classification model. The larger the area, the better the performance.

We will use the method in R-Studio.

## Chapter-5

### Calculations & Analysis:

We have used **R Studio** for all the calculations and analysis. First, we load the data at R Studio. Then the calculations and the analysis are done as follows.

We will use the same procedure for both Male & Female Data.

- **For Female :**

#### **Create Training and Test Samples:**

We split the whole dataset for female into a training set and a testing set. Training set is used to train the model on and testing set is used to test the model on. i.e., we fit the model on the train data and accuracy of fitting is checked on test data. Here 70% of the whole data is taken as train data and remaining 30% data is taken as test data.

Number of rows in train data= 94

Number of rows in test data= 40

#### **❖ Fitting the logistic model**

The multiple logistic regression model is fitted on the train data. The estimate of the coefficients and the p-values are given below.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.85312    2.88169  -2.378 0.017399 *
AGE          0.06147    0.04050   1.518 0.129036
SMOKING      -1.65706    1.18649  -1.397 0.162532
YELLOW_FINGERS 2.03437    1.38720   1.467 0.142504
ANXIETY       3.02208    1.12735   2.681 0.007347 **
PEER_PRESSURE 3.58006    1.05302   3.400 0.000674 ***
FATIGUE       1.70925    1.09915   1.555 0.119930
ALLERGY       4.11480    1.37918   2.984 0.002850 **
WHEEZING      1.13609    0.90390   1.257 0.208795
COUGHING     -1.26531    1.23981  -1.021 0.307460
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients in the output indicate the average change in log odds of class. For example, a one unit increase in **Anxiety** is associated with an average increase of **0.007347** in the log odds of Class.

The p-values in the output gives an idea of how effective each predictor variable is at predicting the probability of Class. Here we can see that **Anxiety, Peer pressure and Allergy** are important predictors since they have low p-values (<0.05) while the others are not nearly as important.

Now with these important covariates we again fit the logistic regression model. And the summary is as follows,

```

coefficients:
              Estimate Std. Error z value Pr(>|z|)
](Intercept)  -1.4613      0.5744  -2.544 0.010954 *
ANXIETY        2.2350      0.7442   3.003 0.002672 **
PEER_PRESSURE  3.1210      0.7723   4.041 5.31e-05 ***
ALLERGY        3.1949      0.9005   3.548 0.000388 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

We can see that here all the predictors are important predictors as the p-values are very low (<0.05).

### ▪ Assessing Model Fit

For logistic regression we compute a metric known as **McFadden's  $R^2$** , which ranges from 0 to just under 1. Values close to 0 indicate that the model has no predictive power. In practice, values over 0.40 indicate that a model fits the data very well.

We can compute McFadden's  $R^2$  for our model using the **p $R^2$**  function from the pscl package. The McFadden's  $R^2$  for the train data is:

**McFadden** = 0.5815674.

This is a high value for McFadden's  $R^2$ , which indicates that our model fits the data very well and has high predictive power.

### ▪ Variable Importance

Higher values indicate more importance.

These results match up nicely with the p-values from the model. Peer Pressure is by far the most important predictor variable, followed by Allergy and then Fatigue.

```

> caret::varImp(model)
              overall
ANXIETY        3.003150
PEER_PRESSURE  4.041418
ALLERGY        3.547784

```

- As a rule of thumb, VIF values above 5 indicate severe multicollinearity. Since none of the predictor variables in our models have a VIF over 5, we can assume that multicollinearity is not an issue in our model.

### • Multicollinearity Checking

#### VIF Values

We calculate the VIF values of each variable for train data in the model to check the multicollinearity.

```

> car::vif(model)
      ANXIETY PEER_PRESSURE      ALLERGY
1.095288      1.101561      1.136863
> |

```

As a rule of thumb, VIF values above 5 indicate severe multicollinearity. Since none of the predictor variables in our models have a VIF over 5, we can assume that multicollinearity is not an issue in our model.

## • Correlation Heatmap of Our Data

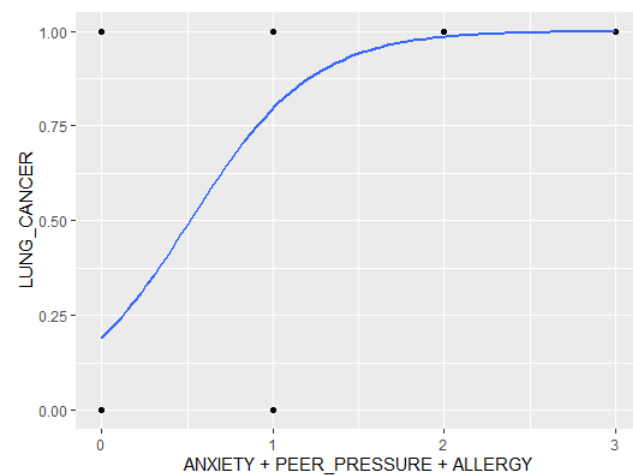
The correlation heat map for our data is as follows,

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
Column 1	1							
Column 2	-0.13422	1						
Column 3	0.005216	0.270844	1					
Column 4	0.005446	0.325646	0.540277	1				
Column 5	0.060496	0.078392	-0.07482	-0.11269	1			
Column 6	-0.02867	-0.17692	-0.05713	-0.00609	-0.08838	1		
Column 7	0.074901	-0.17803	0.091425	0.110216	-0.00363	0.222595	1	
Column 8	0.109586	0.039871	0.319307	0.334961	0.070146	0.32382	0.254112	1

Where Column 1-8 represents Age, Smoking, Yellow Fingers, Anxiety, Fatigue, Allergy and Wheezing respectively.

## ▪ Plotting the Data

We plot our train data by using ggplot, which is used to construct the initial plot object, and is almost always followed by a plus sign (+) to add components to the plot. A logistic model shows **S shaped** curve, called **Sigmoid**. The model fitted on the train data also shows a S shaped curve or Sigmoid curve. From here we can tell that, multiple logistic regression model fits our data very well. The plotted curve is given below,



- **Using the Model to Make Predictions**

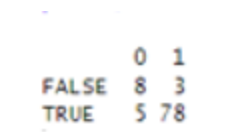
We have fitted the logistic regression model to the train data. Now we will predict the response variable (i.e. - Lung Cancer) for the test data based on the model fitted to the train data. The prediction of the Class is made based on the covariates Smoking Yellow fingers, Anxiety, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, Chest pain .

- **Model Diagnosis**

Now, we analyse how well our model performs on the test data.

- **Confusion Matrix**

We create confusion matrix which shows our predictions compared to the actual Class. The confusion matrix is as follows,



A confusion matrix table with two rows labeled 'FALSE' and 'TRUE' on the left, and two columns labeled '0' and '1' at the top. The values in the cells are: FALSE 0 is 8, FALSE 1 is 3, TRUE 0 is 5, and TRUE 1 is 78.

	0	1
FALSE	8	3
TRUE	5	78

From the confusion matrix we calculate sensitivity, specificity, total misclassification error rate and accuracy for the data.

### **Sensitivity**

We calculate sensitivity from confusion matrix and the value for our trend data is 0.61539 It is high so our model is able to predict the outcomes. So, this particular model turns out to be very good at predicting whether a class is malignant or benign.

### **Specificity**

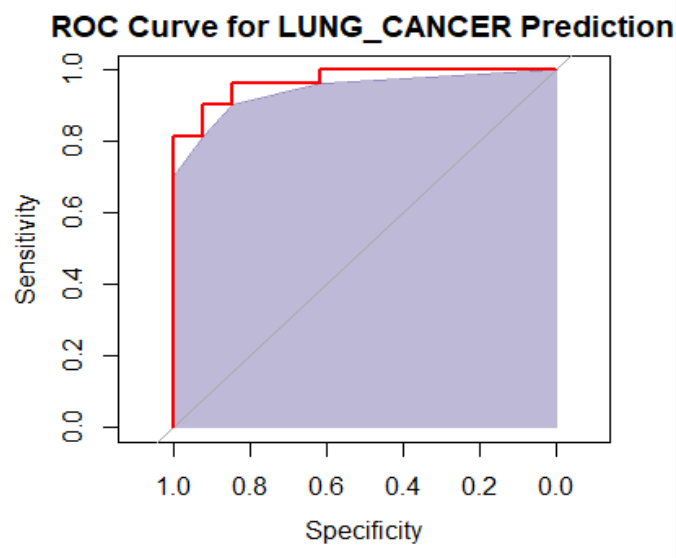
The specificity for our test data is **0.9397690** which indicates that this particular model fits our data very well.

### **Accuracy**

The accuracy for our data is **0.9148**, which is very high. In another way beside total misclassification error rate by accuracy we can say that, the logistic regression model fits the data very well.

### **ROC Curve**

Lastly, we can plot the ROC (Receiver Operating Characteristic) Curve which displays the percentage of true positives predicted by the model as the prediction probability cutoff is lowered from 1 to 0. AUC (the area under the curve) for train data is **0.9435**, which is very high. This indicates that our model does a good job of predicting the Class of an individual. The ROC curve is as follows,



So, This is our ROC Curve for Female Data ..

- **For Male :**

Now we'll do the same work for Male data. We'll check the accuracy of the data and plot ROC Curve...

- **Create Training and Test Samples:**

We split the whole dataset for female into a training set and a testing set. Training set is used to train the model on and testing set is used to test the model on. i.e., we fit the model on the train data and accuracy of fitting is checked on test data. Here 70% of the whole data is taken as train data and remaining 30% data is taken as test data.

Number of rows in train data= 100

Number of rows in test data= 42

- **Fitting the logistic model**

The multiple logistic regression model is fitted on the train data. The estimate of the coefficients and the p-values are given below.



```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.646301   3.014987  -1.209  0.22651
AGE           -0.008315   0.050990  -0.163  0.87046
SMOKING        1.237250   0.718776   1.721  0.08519 .
YELLOW_FINGERS 2.027647   0.915899   2.214  0.02684 *
ANXIETY        1.375365   0.966706   1.423  0.15481
PEER_PRESSURE  1.136815   0.811552   1.401  0.16128
FATIGUE        2.188469   0.873215   2.506  0.01220 *
ALLERGY        3.005754   0.948855   3.168  0.00154 **
WHEEZING       0.850668   0.869447   0.978  0.32788
COUGHING       1.931910   0.970508   1.991  0.04652 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The coefficients in the output indicate the average change in log odds of class. For example, a one unit increase in **Yellow Fingers** is associated with an average increase of **0.02684** in the log odds of Class.

The p-values in the output gives an idea of how effective each predictor variable is at predicting the probability of Class. Here we can see that **Yellow Fingers, Fatigue and Allergy** are important predictors since they have low p-values (<0.05) while the others are not nearly as important.

Now with these important covariates we again fit the logistic regression model. And the summary is as follows,

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8883    0.6394  -1.389  0.164766
YELLOW_FINGERS 1.7467    0.7033   2.483  0.013012 *
FATIGUE       2.0863    0.6880   3.033  0.002425 **
ALLERGY       2.8154    0.7603   3.703  0.000213 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

We can see that here all the predictors are important predictors as the p-values are very low (<0.05).

### ▪ Assessing Model Fit

For logistic regression we compute a metric known as **McFadden's  $R^2$** , which ranges from 0 to just under 1. Values close to 0 indicate that the model has no predictive power. In practice, values over 0.40 indicate that a model fits the data very well.

We can compute McFadden's  $R^2$  for our model using the **p $R^2$**  function from the pscl package. The McFadden's  $R^2$  for the train data is:

- **McFadden** = 0.39953..

This is a high value for McFadden's  $R^2$ , which indicates that our model fits the data very well and has high predictive power.

### ▪ Variable Importance

Higher values indicate more importance.

```

Overall
YELLOW_FINGERS 2.483450
FATIGUE         3.032587
ALLERGY        3.703034

```

These results match up nicely with the p-values from the model Allergy is by far the most important predictor variable, followed by Yellow Fingers and then Fatigue.

- As a rule of thumb, VIF values above 5 indicate severe multicollinearity. Since none of the predictor variables in our models have a VIF over 5, we can assume that multicollinearity is not an issue in our model.

## • Multicollinearity Checking

### VIF Values

We calculate the VIF values of each variable for train data in the model to check the multicollinearity.

```

YELLOW_FINGERS 1.314770
FATIGUE        1.219443
ALLERGY        1.186989

```

As a rule of thumb, VIF values above 5 indicate severe multicollinearity. Since none of the predictor variables in our models have a VIF over 5, we can assume that multicollinearity is not an issue in our model.

## • Correlation Heatmap of Our Data

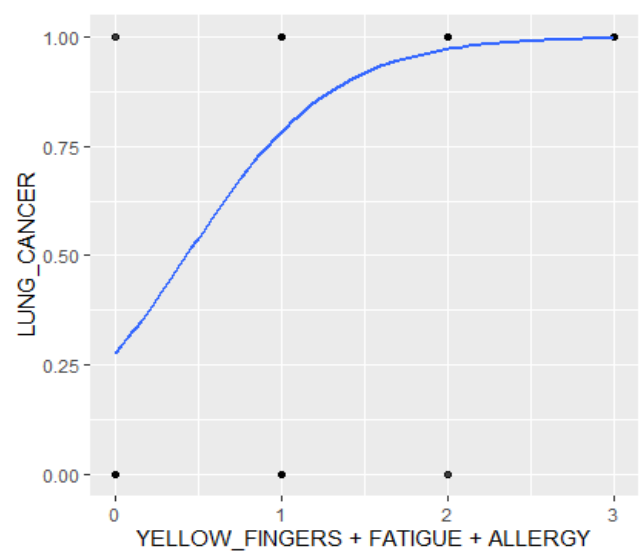
The correlation heat map for our data is as follows,

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
Column 1	1							
Column 2	-0.00414	1						
Column 3	0.043271	-0.26462	1					
Column 4	0.098764	0.005668	0.549874	1				
Column 5	0.001049	-0.15092	-0.16379	-0.28314	1			
Column 6	0.118606	0.100105	-0.17832	-0.26969	0.085332	1		
Column 7	0.032042	-0.12931	-0.14951	-0.41827	0.330659	0.080269	1	
Column 8	0.104247	0.025258	0.092966	-0.03588	0.253631	0.336718	0.235069	1

Where Column 1-8 represents Age, Smoking, Yellow Fingers, Anxiety, Fatigue, Allergy and wheezing respectively.

## ▪ Plotting the Data

We plot our train data by using ggplot, which is used to construct the initial plot object, and is almost always followed by a plus sign (+) to add components to the plot. A logistic model shows **S shaped** curve, called **Sigmoid**. The model fitted on the train data also shows a S shaped curve or Sigmoid curve. From here we can tell that, multiple logistic regression model fits our data very well. The plotted curve is given below,



- **Using the Model to Make Predictions**

We have fitted the logistic regression model to the train data. Now we will predict the response variable (i.e. - Lung Cancer) for the test data based on the model fitted to the train data. The prediction of the Class is made based on the covariates Smoking Yellow fingers, Anxiety, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, Chest pain .

- **Model Diagnosis**

Now, we analyse how well our model performs on the test data.

- **Confusion Matrix**

We create confusion matrix which shows our predictions compared to the actual Class. The confusion matrix is as follows,

	0	1
FALSE	4	1
TRUE	8	87

From the confusion matrix we calculate sensitivity, specificity, total misclassification error rate and accuracy for the data.

## Sensitivity

We calculate sensitivity from confusion matrix and the value for our trend data is 0.5 It is high so our model is able to predict the outcomes. So, this particular model turns out to be very good at predicting whether a class is malignant or benign.

## Specificity

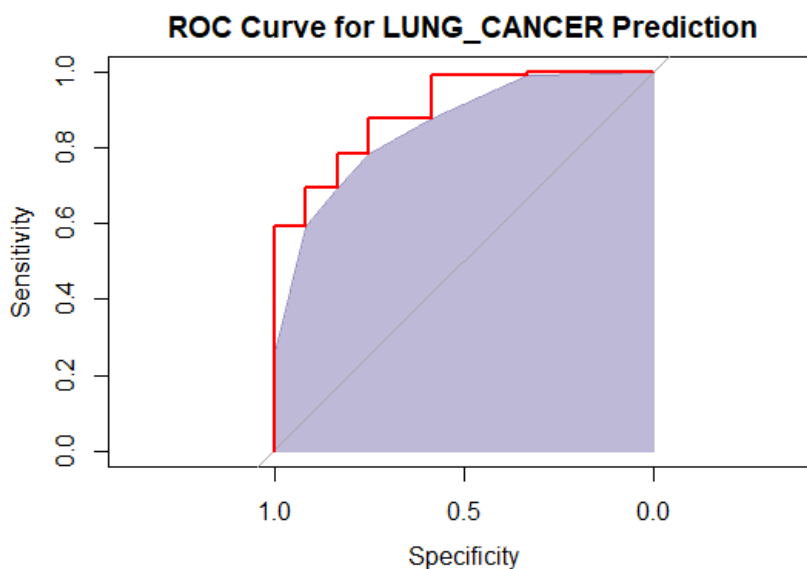
The specificity for our test data is **0.9456521739** which indicates that this particular model fits our data very well.

## Accuracy

The accuracy for our data is **0.91**, which is very high. In another way beside total misclassification error rate by accuracy we can say that, the logistic regression model fits the data very well.

## ROC Curve

Lastly, we can plot the ROC (Receiver Operating Characteristic) Curve which displays the percentage of true positives predicted by the model as the prediction probability cutoff is lowered from 1 to 0. AUC (the area under the curve) for train data is **0.8532**, which is very high. This indicates that our model does a good job of predicting the Class of an individual. The ROC curve is as follows,



So, This is our ROC Curve for Male Data.

## **Chapter: 6**

### **CONCLUSION:**

I've divided the female data set in 94 rows of train data and 40 rows of test data and the male data set in 100 rows of train data and 42 rows of test data. After calculating the p-values of all the factors we can see that **Anxiety, Peer pressure and Allergy** are important predictors for Female data and **Yellow Fingers, Fatigue and Allergy** are important predictors for male data. **For both cases our McFadden's  $R^2$  value is consistent** and none of the predictor variables have a VIF value over 5 so we can assume that **multicollinearity is not an issue in our model**. Due to the **high specificity rate** (for both male and female data) we can say that the model we chose, fits our data very well. For Female data, the area under the ROC curve is 0.9435 and for Male data, the area under the ROC curve is 0.8532. This indicates **the quality of classification model is good**.

From this report we can conclude that Lung Cancer Data of Female Patients will fit better in our model than Male Patient, i.e. Female Data will fit **94 % accurately** & Male Data will fit **85.32 % accurately** in our model.

## **Chapter 7**

### **REFERENCE:**

- Fundamental of Statistics Vol -1 & Vol – 2 by Gun Gupta Dasgupta
- An Introduction to Categorical Data Analysis by Alan Agresti
- Logistic Regression A Self-Learning Text by David G. Kleinbaum, Mitchel Klein.

## **Chapter: 8**

### **ACKNOWLEDGEMENT:**

I would like to express my special thanks of gratitude to my teacher and mentor of this project Prof. Dithi Bhattacharya for her constant support and guidance. I also thank to all teacher of statistics department who gave me knowledge which helped me to complete this project. Secondly I want to thank University of Calcutta for giving me such opportunity.

I am overwhelmed in all humbleness and gratefulness to acknowledge my depth to all my friends who have helped and supported me in the whole project. And lastly I want to thank my parents who also guided me giving precious advice in making this project.

----- THANK YOU -----

## **Chapter: 9**

### **Annexure:**

- For Female :

#Number of Rows and Column

```
nrow(Lung)
```

```
ncol(Lung)
```

#Dividing in Train & Test data

```
library(caret)
```

```
library(ggplot2)
```

```
library(lattice)
```

```
set.seed(123)
```

```
trainIndex<-createDataPartition( Lung$ Lung_Cancer,p=0.7,list=FALSE)
```

```
train<-Cancer[trainIndex,]
```

```
test<-Cancer[-trainIndex,]
```

```
summary(train)
```

```
nrow(train)
```

```
nrow(test)
```

#Model fitting and checking

```
data=read.table("lung")
```

```
model <-
```

```
glm(LUNG_CANCER~AGE+SMOKING+ YELLOW_FINGERS+ANXIETY+FATIGUE+  
ALLERGY+WHEEZING, family="binomial", data=lung)
```

```
summary(model)
```

```
model <- glm(LUNG_CANCER~ANXIETY+PEER_PRESSURE+ALLERGY,  
family="binomial", data=lung)
```



```
summary(model)
```

```
library(psc1)
```

```
psc1::pR2(model)["McFadden"]
```

```
library(caret)
```

```
caret::varImp(model)
```

```
library(car)
```

```
car::vif(model)
```

## #Plotting

```
library(caret)
```

```
library(ggplot2)
```

```
library(lattice)
```

```
#plot logistic regression curve
```

```
ggplot(model, aes(x=ANXIETY+PEER_PRESSURE+ALLERGY, y=LUNG_CANCER.)) +
```

```
  geom_point(alpha=.5) +
```

```
  stat_smooth(method="glm", se=FALSE, method.args = list(family=binomial))
```

## #Prediction

```
predicted<-predict(model,test,type="response")
```

```
summary(predicted)
```

## #Confusion Matrix

```
confusion_matrix<-table(predicted>0.5,test$Lung_Cancer)
```

```
confusion_matrix
```

## #Accuracy

```
accuracy<-sum(diag(confusion_matrix))/sum(confusion_matrix)
```

```
print(accuracy)
```

## #Sensitivity & Specificity

```
sensitivity <- confusion_matrix[2, 2] / (confusion_matrix[2, 2] + confusion_matrix[2, 1])
```

```
sensitivity
```

```
Specificity <- confusion_matrix[1, 1] / (confusion_matrix[1, 1] + confusion_matrix[1, 2])
```

```
specificity
```

```
#AUC & ROC Curve
```

```
install.packages("pROC")
```

```
library(pROC)
```

```
roc_obj<-roc(test$Lung_Cancer,predicted)
```

```
auc<-auc(roc_obj)
```

```
print(auc)
```

```
plot(roc_obj,col="blue",type="shape",auc.polygon=TRUE,auc.polygon.col=rgb(0.35,0.31,0.61,alpha=0.4),auc.polygon.border=rgb(0.35,0.31,0.61,0.4),main="ROC Curve for Type of Lung Cancer Prediction")
```

- For Male :

```
#Number of Rows and Column
```

```
nrow(Lungm)
```

```
ncol(Lungm)
```

```
#Dividing in Train & Test data
```

```
library(caret)
```

```
library(ggplot2)
```

```
library(lattice)
```

```
set.seed(123)
```

```
trainIndex<-createDataPartition( Lungm$ Lung_Cancer,p=0.7,list=FALSE)
```

```
train<-Cancer[trainIndex,]
```

```
test<-Cancer[-trainIndex,]
```

```
summary(train)
```

```
nrow(train)
```

```
nrow(test)
```

## #Model fitting and checking

```
data=read.table("lungm")
```

```
model <-
```

```
glm(LUNG_CANCER~AGE+SMOKING+ YELLOW_FINGERS+ANXIETY+FATIGUE+  
ALLERGY+WHEEZING, family="binomial", data=lungm)
```

```
summary(model)
```

```
model <- glm(LUNG_CANCER~ANXIETY+PEER_PRESSURE+ALLERGY,  
family="binomial", data=lung)
```

```
summary(model)
```

```
library(psc1)
```

```
psc1::pR2(model)["McFadden"]
```

```
library(caret)
```

```
caret::varImp(model)
```

```
library(car)
```

```
car::vif(model)
```

## #Plotting

```
library(caret)
```

```
library(ggplot2)
```

```
library(lattice)
```

```
#plot logistic regression curve
```

```
ggplot(model, aes(x=YELLOW_FINGERS+FATIGUE +ALLERGY, y=LUNG_CANCER.))  
+ geom_point(alpha=.5) + stat_smooth(method="glm", se=FALSE, method.args =  
list(family=binomial))
```

## #Prediction

```
predicted<-predict(model,test,type="response")
```

```
summary(predicted)
```

### #Confusion Matrix

```
confusion_matrix<-table(predicted>0.5,test$Lung_Cancer)

confusion_matrix
```

### #Accuracy

```
accuracy<-sum(diag(confusion_matrix))/sum(confusion_matrix)

print(accuracy)
```

### #Sensitivity & Specificity

```
sensitivity <- confusion_matrix[2, 2] / (confusion_matrix[2, 2] + confusion_matrix[2, 1])

sensitivity

Specificity <- confusion_matrix[1, 1] / (confusion_matrix[1, 1] + confusion_matrix[1, 2])

specificity
```

### #AUC & ROC Curve

```
install.packages("pROC")

library(pROC)

roc_obj<-roc(test$Lung_Cancer,predicted)

auc<-auc(roc_obj)

print(auc)
```

```
plot(roc_obj,col="blue",type="shape",auc.polygon=TRUE,auc.polygon.col=rgb(0.35,0.31,0.61,alpha=0.4),auc.polygon.border=rgb(0.35,0.31,0.61,0.4),main="ROC Curve for Type of Lung Cancer Prediction")
```