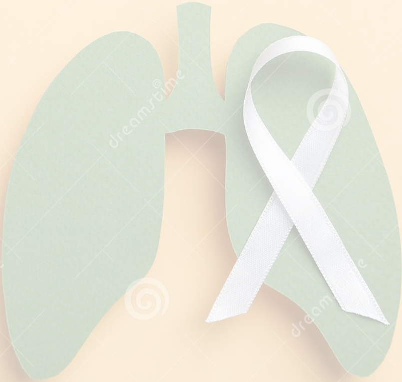
Statistical Analysis and Predicting Lung Cancer

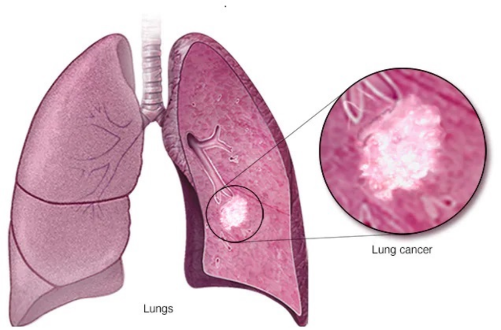
Bidisha Bhandari

Visva Bharati University

May 23, 2023



INTRODUCTION



- **Lung cancer begins in the lungs and may spread to lymph nodes or other organs in the body or vice-versa.**
- **Lung cancers usually are grouped into two main types called small cell and non-small cell (including adenocarcinoma and squamous cell carcinoma).**

START

About the dataset

- **The dataset contains the 309 responses irrespective of age and sex regarding to the symptoms of Lung Cancer.**
- **Pure categorical dataset**
- **https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer**

RAW DATA

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC_DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL_CONSUMING	COUGHING	SHORT
M	69	0	1	1	0	0	1	0	1	1	1	
M	74	1	0	0	0	1	1	1	0	0	0	
F	59	0	0	0	1	0	1	0	1	0	1	
M	63	1	1	1	0	0	0	0	0	1	0	
F	63	0	1	0	0	0	0	0	1	0	1	
F	75	0	1	0	0	1	1	1	1	0	1	
M	51	1	0	0	0	0	1	0	1	1	1	
F	50	1	1	1	1	0	1	1	0	0	0	
F	68	1	0	1	0	0	1	0	0	0	0	
M	53	1	1	1	1	1	0	1	0	1	0	
F	60	1	1	1	1	1	1	0	1	0	1	
M	71	0	0	0	0	1	1	1	1	1	1	
F	60	1	0	0	0	0	1	0	0	0	0	
M	58	1	0	0	0	0	1	1	1	1	1	
M	69	1	0	0	0	0	0	1	1	1	1	
F	48	0	1	1	1	1	1	1	1	0	1	
M	75	1	0	0	0	1	0	1	1	1	1	
M	57	1	1	1	1	1	0	0	0	1	0	
F	68	1	1	1	1	1	1	0	0	0	1	
F	60	0	0	0	0	1	1	0	0	0	0	
F	44	1	1	1	1	1	1	0	0	0	0	
F	64	0	1	1	1	0	0	1	1	0	1	
F	10	1	0	0	0	1	1	1	0	0	0	

图 1: RAW DATA

1. INTRODUCTION
   - About the dataset
   - DATA TYPE
2. OBJECTIVES
3. METHODOLOGY
4. CONCLUSION
5. AREA OF FOCUS
6. REFERENCES

Data type

- This data is purely categorical. So for this kind of survey data one can follow this project.
- This project shows which factors are more injuries for lung cancer that will spread a mass awareness.
- Visualization is most important for this kind of data. That' s the basic reason to give as much plot as possible and have to give a decision tree to conclude all them at one glance.

Risk Factors of Lung Cancer

- **Smoking**
- **Yellow fingers**
- **Anxiety**
- **Peer pressure**
- **Chronic diseases**
- **Fatigue**
- **Allergy**
- **Wheezing**
- **Alcohol consuming**
- **Coughing**
- **Shortness of breath**
- **Swallowing difficulty**
- **Chest pain**

RISK FACTORS


RISK
FACTORS

OBJECTIVES

- **Primary aim** To find the chance of getting lung cancer by seeing the symptoms.
- **Secondary aim**
  1. To spread self-awareness.
  2. To decrease the mortality caused by lung cancer
  3. To find out the chance how many among the sample population will get lung cancer in future.

PIE-CHART

Syntax for Pie chart

```
pie(x, labels, main, col)
```

PIE-CHART

Syntax for Pie chart

pie(x, labels, main, col)

Gender distribution suffering lung cancer

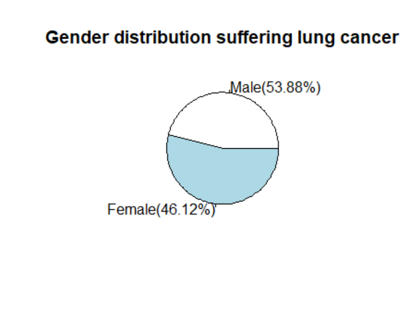


图 2: Gender distribution suffering lung cancer

PIE-CHART

Syntax for Pie chart

pie(x, labels, main, col)

Gender distribution suffering lung cancer

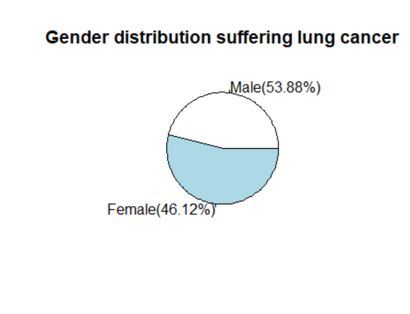


图 2: Gender distribution suffering lung cancer

Proportion of ’yes’ responses of regarding factors

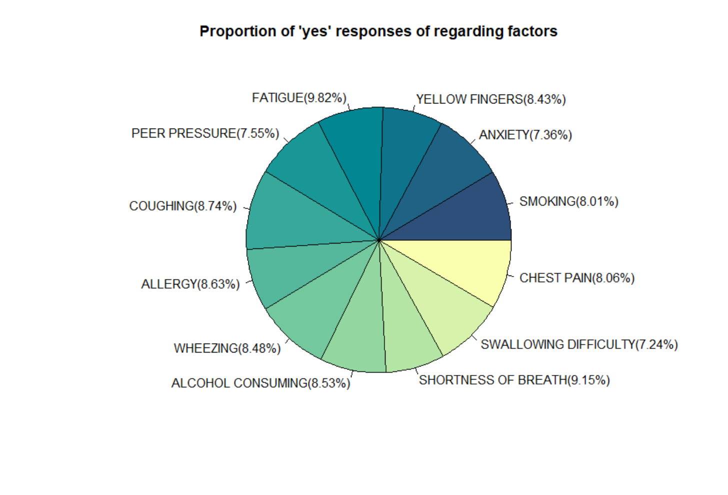


图 3: Proportion of ’yes’ responses of regarding factors

HISTOGRAM

Syntax for Histogram

```
hist(v,main,xlab,xlim,ylim,breaks,col,border)
```

HISTOGRAM

Syntax for Histogram

```
hist(v,main,xlab,xlim,ylim,breaks,col,border)
```

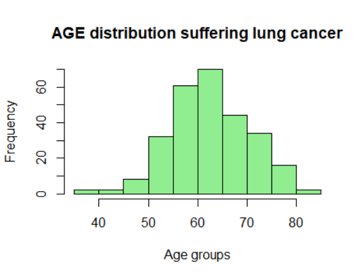


图 4: Age distribution suffering lung cancer

INTERPRETATION 60-65 age group is in more risk than others.

BARPLOT

Syntax for Barplot

```
barplot(H, xlab, ylab, main,col)
```

BARPLOT

Syntax for Barplot

barplot(H, xlab, ylab, main,col)

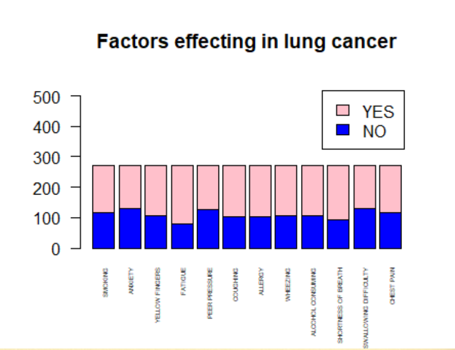


图 5: Factors effecting in lung cancer

HEATMAP

- Correlation between each and every factors of our data.

HEATMAP

- Correlation between each and every factors of our data.

图 6: Heatmap

LOGISTIC REGRESSION

Logistic Regression Model

$$\boxed{Logit(p_i) = 1/(1 + exp(-p_i))}$$

$$\boxed{ln(p_i/(1 - p_i)) = \beta_0 + \beta_1 * X_1 + \ldots + B_k * X_k}$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE).

```
> summary(logistic_model)

Call:
glm(formula = LUNG_CANCER ~ ., family = binomial(), data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.46835   0.00214   0.01849   0.13654   2.43803

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -15.41386    4.21530  -3.657 0.000256 ***
GENDERM                -0.47702    1.01063  -0.472 0.636923
AGE                     0.06718    0.03768   1.783 0.074585 .
SMOKING                 4.58089    1.64650   2.782 0.005399 **
YELLOW_FINGERS          2.68033    1.17600   2.279 0.022655 *
ANXIETY                 0.11886    1.19701   0.099 0.920899
PEER_PRESSURE           2.18619    1.09357   1.999 0.045595 *
CHRONIC_DISEASE         5.50127    1.82045   3.022 0.002512 **
FATIGUE                 4.69539    1.47570   3.182 0.001464 **
ALLERGY                 0.65779    1.15642   0.569 0.569482
WHEEZING                1.39511    1.29694   1.076 0.282064
ALCOHOL_CONSUMING       2.60999    1.30705   1.997 0.045841 *
COUGHING                3.55012    1.68159   2.111 0.034758 *
SHORTNESS_OF_BREATH    -1.19793    1.31663  -0.910 0.362907
SWALLOWING_DIFFICULTY   4.81894    2.01262   2.394 0.016650 *
`CHEST-PAIN`            1.79482    1.05613   1.699 0.089237 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.382  on 215  degrees of freedom
Residual deviance:  47.568  on 200  degrees of freedom
AIC: 79.568

Number of Fisher Scoring iterations: 9
```

INTERPRETATION

- Each one-unit change in yellow fingers will increase the log odds of getting lung cancer by 2.68, and its p-value indicates that it is somewhat significant in determining the lung cancer.
- Each unit increase in peer pressure increases the log odds of getting lung cancer by 2.18 and p-value indicates that it is somewhat significant in determining the lung cancer.
- Similarly we can interpret for other factors also

PROBISTIC REGRESSION

Probit regression model

$\boxed{Pr(Y=1|X)=\phi(\beta_0+\beta_1 X)}$

Where, is the cumulative normal distribution function and $z=\beta_0+\beta_1 X$ is the "z-value" or "z-index" of the probit model.

```
> summary(probistic_model)

Call:
glm(formula = LUNG_CANCER ~ ., family = gaussian(), data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.80861  -0.11966   0.03153   0.16062   0.66488

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.118438   0.151732  -0.781 0.435977
GENDERM                0.011986   0.042728   0.281 0.779364
AGE                    0.002582   0.002217   1.165 0.245395
SMOKING                0.105674   0.037756   2.799 0.005631 **
YELLOW_FINGERS         0.138775   0.046687   2.972 0.003317 **
ANXIETY                0.081986   0.050415   1.626 0.105478
PEER_PRESSURE          0.105048   0.044316   2.370 0.018719 *
CHRONIC_DISEASE        0.109338   0.038346   2.851 0.004810 **
FATIGUE                0.193859   0.045167   4.292 2.76e-05 ***
ALLERGY                0.140062   0.039972   3.504 0.000565 ***
WHEEZING               0.073410   0.040213   1.826 0.069410 .
ALCOHOL_CONSUMING      0.215034   0.045987   4.676 5.37e-06 ***
COUGHING               0.096919   0.043737   2.216 0.027824 *
SHORTNESS_OF_BREATH    0.055034   0.046484   1.184 0.237852
SWALLOWING_DIFFICULTY  0.102112   0.045607   2.239 0.026261 *
`CHEST-PAIN`           0.049770   0.039526   1.259 0.209439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06683455)

    Null deviance: 25.106  on 215  degrees of freedom
Residual deviance: 13.367  on 200  degrees of freedom
AIC: 45.962

Number of Fisher Scoring iterations: 2
```

INTERPRETATION

- Each one-unit change in coughing will increase the z score of getting lung cancer by 0.09, and its p-value indicates that it is somewhat significant in determining the lung cancer.
- Each unit increase in peer pressure increases the z score of getting lung cancer by 0.105 and p-value indicates that it is somewhat significant in determining the lung cancer.
- Each unit increase in swallowing difficulty increases the z score of getting lung cancer by 0.102 and p-value indicates that it is somewhat significant in determining the lung cancer.

POISSON REGRESSION

Poisson regression model

If $\mathbf{x} \in \mathbb{R}^n$ is a vector of independent variables, then the model takes the form

$$\boxed{\log(\mathrm{E}(Y \mid \mathbf{x})) = \alpha + \beta' \mathbf{x}}$$

POISSON REGRESSION

Poisson regression model

If $\mathbf{x} \in \mathbb{R}^n$ is a vector of independent variables, then the model takes the form

$$\boxed{\log(\mathrm{E}(Y \mid \mathbf{x})) = \alpha + \beta' \mathbf{x}}$$

where

$\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$. Sometimes this is written more compactly as

$\log(\mathrm{E}(Y \mid \mathbf{x})) = \boldsymbol{\theta}' \mathbf{x}$,

where $\mathbf{x}$ is now an (n + 1)-dimensional vector consisting of n independent variables concatenated to the number one. Here

θ is simply α concatenated to β

Thus, when given a Poisson regression model θ and an input vector

$\mathbf{x}$, the predicted mean of the associated Poisson distribution is given by $\mathrm{E}(Y \mid \mathbf{x}) = e^{\boldsymbol{\theta}' \mathbf{x}}$

If Y_i are independent observations with corresponding values $\mathbf{x}_i$ of the predictor variables, then θ can be estimated by maximum likelihood.

```
> summary(Poisson_Regression)

Call:
glm(formula = LUNG_CANCER ~ ., family = poisson(), data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.25937  -0.13523   0.04693   0.19814   0.75471

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.437019   0.666699  -2.155   0.0311 *
GENDERM                0.020595   0.179859   0.115   0.9088
AGE                    0.003033   0.009368   0.324   0.7461
SMOKING                0.135289   0.159750   0.847   0.3971
YELLOW_FINGERS         0.181304   0.197778   0.917   0.3593
ANXIETY                0.114933   0.212174   0.542   0.5880
PEER_PRESSURE          0.128508   0.191831   0.670   0.5029
CHRONIC_DISEASE        0.129527   0.159998   0.810   0.4182
FATIGUE                0.251490   0.202072   1.245   0.2133
ALLERGY                0.180764   0.168859   1.071   0.2844
WHEEZING               0.090697   0.165784   0.547   0.5843
ALCOHOL_CONSUMING      0.293513   0.197234   1.488   0.1367
COUGHING               0.122862   0.181749   0.676   0.4990
SHORTNESS_OF_BREATH    0.077808   0.195013   0.399   0.6899
SWALLOWING_DIFFICULTY  0.130352   0.190290   0.685   0.4933
`CHEST-PAIN`           0.048356   0.167956   0.288   0.7734
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 53.920  on 215  degrees of freedom
Residual deviance: 39.409  on 200  degrees of freedom
AIC: 445.41

Number of Fisher Scoring iterations: 5
```

COMPARISON BETWEEN REGRESSION MODELS

$$\boxed{AIC = 2K – 2ln(L)}$$

The model with the lowest AIC offers the best fit.

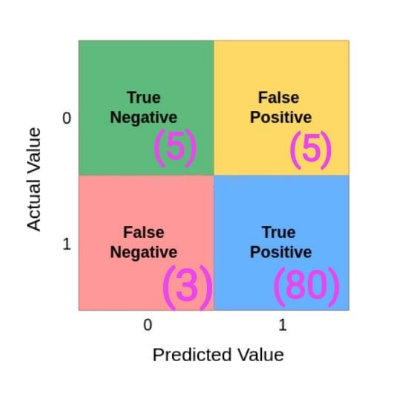
COMPARISON BETWEEN REGRESSION MODELS

$$\boxed{AIC = 2K - 2ln(L)}$$

The model with the lowest AIC offers the best fit.

AIC of **Probistic** model(AIC=45.962) is ***lower*** than logistic model **(AIC=79.568)** and regression with poisson family**(AIC= 445.41)**. So we may conclude that ***Probistic model is better fit*** for this model.

CONFUSION MATRIX



$\boxed{\text{Accuracy} = 0.914}$

INTERPRETATION OF CONFUSION MATRIX

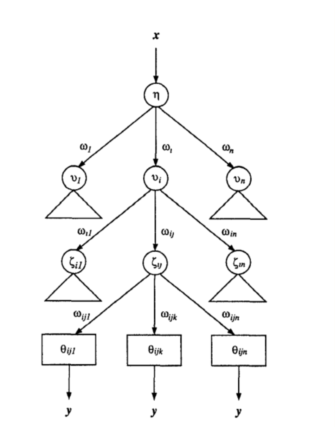
1. Positive class are 80+3=83 Negative class, which is (5+5=)10
2. Correct classifications are the diagonal elements of the matrix 80 for the positive class and 5 for the negative class.
3. 3 samples (bottom-left box) were expected to be of the positive class but were classified as the "negative" by the model 5 samples (top-right box) were expected to be of negative class but were classified as "positive" by the model

Probabilistic model of a decision tree

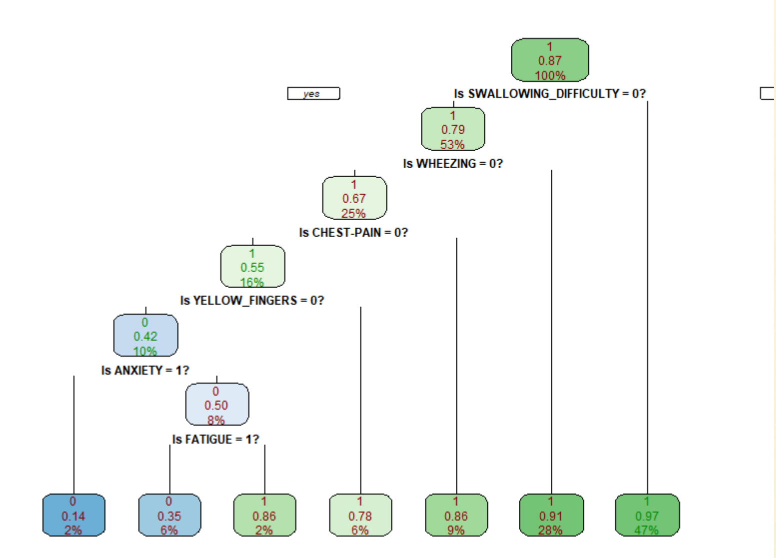
- A probabilistic model of a decision tree involves a sequence of probabilistic decisions, each conditional on the input z and conditional on previous decisions.

Probabilistic model of a decision tree

- A probabilistic model of a decision tree involves a sequence of probabilistic decisions, each conditional on the input z and conditional on previous decisions.



DECISION TREE BASED ON HOSPITAL DATA



Interpret the probability of getting lung cancer I

1. P(53 % of the total population | no swallowing difficulty) = 0.79
2. P(47 % of the total population | swallowing difficulty) = 0.97
3. P(25 % of the (1) population | no swallowing difficulty, no wheezing) = 0.67
4. P(28 % of the (1) population | no swallowing difficulty,wheezing) = 0.91
5. P(16 % of the (3) population | no swallowing difficulty,no wheezing,no chest pain) = 0.67
6. P(9 % of the (3) population | no swallowing difficulty,no wheezing, chest pain) = 0.86 item P(10 % of the (5) population | no swallowing difficulty,no wheezing, no chest pain,no yellow fingers) = 0.42
7. P(6 % of the (5) population | no swallowing difficulty,no wheezing, no chest pain,yellow fingers) = 0.78
8. P(8 % of the (7) population | no swallowing difficulty,no wheezing, no chest pain,no yellow fingers,no anxiety) = 0.50
9. P(2 % of the (7) population | no swallowing difficulty,no wheezing, no chest pain,no yellow fingers, anxiety) = 0.14

Interpret the probability of getting lung cancer II

10. P(6 % of the (9) population | no swallowing difficulty,no wheezing, no chest pain,no yellow fingers,no anxiety,fatigue) = 0.35
11. P(2 % of the (9) population | no swallowing difficulty,no wheezing, no chest pain,no yellow fingers,no anxiety,no fatigue) = 0.14

ANOTHER REPRESENTATION OF DECISION TREE

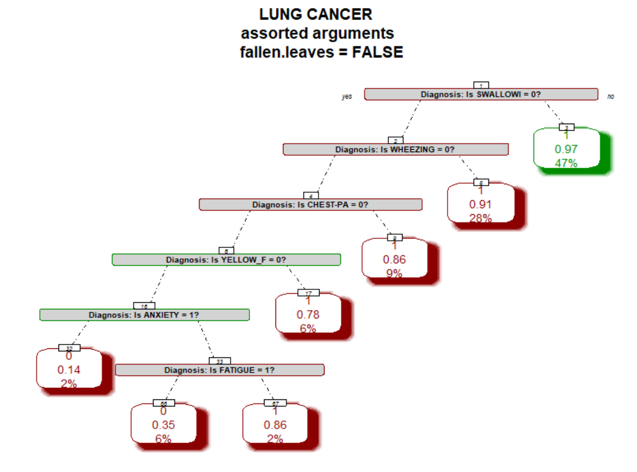


图 9: Another representation of decision tree

CONCLUSION

- **If the medical reports of the sample (the population taken for survey) are available or if the symptoms of the sample population are observed, then this project will be useful to find out the chance how many among the sample population will get lung cancer in future.**
- **Thus the project can be a medical forecast or maybe a medical support for the population on whom the survey is been conducted.**


CONCLUSION

AREA OF FOCUS

OPPORTUNITY

- This project is a great opportunity for those who wants to work with Hospital data . This will complete the whole survey at once.

- With the help of this project, one can estimates his/her own situation (or position) in the risk of getting lung cancer.


References

REFERENCES

- *https : //www.cdc.gov/cancer/lung/basic$_i$nfo/what − is − lung − cancer.htm*
- *https : //www.sciencedirect.com*
- *https : //www.guru99.com*
- *https : //www.educba.com*
- *https : //www.tutorialspoint.com*
- *https : //www.wikipedia.org/*

Books

Analysis of Categorical Data with R
Book by Christopher R. Bilder and Thomas M. Loughin


Thank
You