IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 12, December 2019, ||Series -I || PP 52-55

Diagnosis System for Lung Cancer using R Programming

Prasenjit Nath
Assistant Professor B.Voc(IT)
B.N.College, Dhubri, Assam, India
Received 10 December 2019; Accepted 25 December 2019

Abstract: Lung cancer becomes a major thread in North-East India, with a population of approximately 5 corer records 45,000 new cancer cases annually, 70% of which are reported at late stage resulting in a high mortality rate of 50%. The delayed diagnosis is due to a lack of awareness of cancer symptoms, poor access to affordable care and other psychological factors, like fear and fatalism. This study aims to analysis the lung cancer data and tries to find out relationship between chest pain of smoker and change of lung cancer. For this research paper Data Mining algorithm by using R Programming is used. This paper will analyze the lung cancer data which is taken from relationship between chest pain of smoker and change of lung cancer on cancer patient data sets taken from data world team.
Keywords: Data mining, R-Programming, medical diagnosis, Lung Cancer, North East India, Assam.

I. INTRODUCTION:

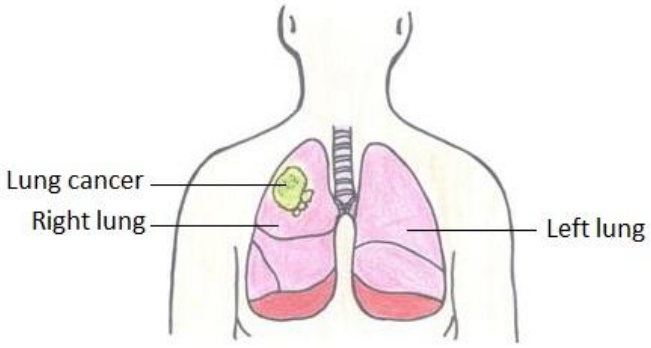
The lungs are a pair of sponge-like cone shaped organs in the chest. These are part of our respiratory system. The left lung is smaller because the heart occupies space on left side. The lungs are slightly different on each side; Right lung has three lobes, whereas the left lung has two lobes. Lungs are covered by a thin covering called 'pleura' which protects and helps lungs move back and forth as they expand and contract during breathing. A thin, dome-shaped muscle below the lungs called 'diaphragm' separates the chest from the abdomen. The diaphragm moves up and down during breathing forcing air in and out of the lungs.

Main function of the lungs is to exchange gases between the air we breathe and the blood. When we breathe in (inhale), oxygen enters into the body through the lungs and when we breathe out (exhale) carbon dioxide is sent out of the body.

Air enters the lungs through nose or mouth via windpipe (trachea) which divides into two airways going into right and left lungs. These airways are called 'bronchi (singular, bronchus). Inside each lung the bronchus divides into smaller tubes, the 'secondary bronchi' which further subdivide into smaller branches called bronchioles. At the end of the bronchioles are tiny air sacs known as 'alveoli'. Many tiny blood vessels that run through these alveoli perform the function of exchange of gases.

What is Lung Cancer?

Lung cancer is a type of cancer that arises in the lungs. It may spread to lymph nodes or other organs in the body, such as the brain. Lung cancers usually are grouped into two main types, non-small cell lung cancer (NSCLC) and small cell lung cancers (SCLC) based on appearance of tumor cells under the microscope. Non-small cell lung cancer (75-80 % of cases) is more common than small cell lung cancer (15-20%) [1].



The number of cancer cases is increasing by the day in the state, a study conducted by B Barooah Cancer Hospital (BBCH) revealed. Assam is the worst-affected state in the northeast. In 2011-12, the maximum number of cancer cases in the hospital has been reported from Kamrup, Nagaon and Dhubri districts.

II. ORIGIN OF RESEARCH PROBLEM:

The North-East Region (NER) of India, with a population of approximately 5 corer records 45,000 new cancer cases annually, 70% of which are reported at late stage resulting in a high mortality rate of 50%. The delayed diagnosis is due to a lack of awareness of cancer symptoms, poor access to affordable care and other psychological factors, like fear and fatalism. It has been proved that communication related to health not only substantially affects perception and behaviour related to health but also linked to health seeking behaviour. Approximately 50% of cancers are preventable by controlling the modifiable risk factors such as tobacco, alcohol, unhealthy diet and physical inactivity, amongst others.

The latest estimates predict 15 lakhs new cases of cancer each year in India with breast, lung and cervical cancers topping the list. One world-wide initiative to quantify cancer as Global Burden of Disease suggests that for some nations, like the US and the UK, some cancer rates (e.g. breast) are beginning to decline. Unfortunately, India's trends show no sign of abetting [2].

The main aim of this study is to develop an expert system which can detect and diagnosis the lung cancer of a patient from the inputs of patient's data and so that can prevent spreading the cancer into the next step. This type of facility at present is not available in North-East India.

III. DATA AND METHODOLOGY:

i. Data and sample

The study is based on secondary data. The data for this study was collected from data world which is a very good site for finding data sets for practicing data science.

ii. The study is basically design to achieve quantative results. Statistical technique like regression analysis has been applied to describe the data using python along with the visualization of data.

iii. The relation between the levels of Lung Cancer has been measured in terms of Smoking, Drinking and age etc.

IV. REVIEW OF LITERATURE:

Dr. Amal Chandra Kataki, the director of B. Borooah Cancer Institute (BBCI), guwahati, Assam stated that out of 4.5 corer population in the north East Region, every year 39,635 new cancer cases are detected, out of which Assam alone contributes to 29,962 patients[3].

Experimental set-up:

At present time millions of data is generated per second and we need different tools that can be utilize to handle these huge amount of data as well as to apply different data mining algorithms along with data visualization in a very quick and effective ways.

R is programming language is a free software environment for statistical data analysis and visualization which is widely used in statisticians and data manipulation and analysis. A lots of visualization options are available in R programming. Using logistic regression in R one can not only explore a data set, but also fit the logistic regression models using the glm() function in R, evaluate the results which is an excellent analysis technique.

```
df <- read.csv("CancerPatient.csv")
new_obs=data.frame(df$Age,df$Gender,df$Obesity,df$Alcohol,df$Smoking,df$Coughing.of.Blood,df$Dry.Cough,df$Level)
lrfit<-glm(formula=cbind(Alcohol,Smoking)~Age+Obesity+Gender+Coughing.of.Blood+
Dry.Cough+ Level, family = 'binomial', data = df)
```

Data Visualization is perhaps the fastest and most useful way to summarize and learn more about use data. One can start by exploring the numeric variables individually. Summary () returns the estimate, standard errors, Z-score, and p-values on each of the coefficients.

V. RESULTS AND DISCUSSION:

Summary () function returns the estimate, standard errors, Z-score, and P-values on each of the coefficients. Look like none of the coefficients are significant here. It also gives you the null deviance (the deviance just for the mean) and the residual deviance (the deviance for the model with all the predictors).

Then assign the result of predict () of glm.prods, with type equals to response. This will make predictions on the training data that one use to fit the model and give me a vector of fitted probabilities.

```
Coefficients:
    (Intercept)            Age        Obesity         Gender
     -0.7333487      0.0008195      0.0368609      0.0380460
Coughing.of.Blood      Dry.Cough       Levellow     LevelMedium
```

0.0405178 0.0507480 -0.0104920 0.6688713

Degrees of Freedom: 999 Total (i.e. Null); 992 Residual
Null Deviance: 773.8
Residual Deviance: 587.9 AIC: 2999

Using summary () function I get following result:

Deviance Residuals:
Min 1Q Median 3Q Max
-1.8804 -0.4728 -0.1084 0.3954 1.8525

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.7333487 0.2066852 -3.548 0.000388 ***
Age 0.0008195 0.0018375 0.446 0.655600
Obesity 0.0368609 0.0213311 1.728 0.083982 .
Gender 0.0380460 0.0483326 0.787 0.431182
Coughing.of.Blood 0.0405178 0.0182551 2.220 0.026450 *
Dry.Cough 0.0507480 0.0118113 4.297 1.73e-05 ***
Levellow -0.0104920 0.1149532 -0.091 0.927277
LevelMedium 0.6688713 0.0840238 7.960 1.71e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

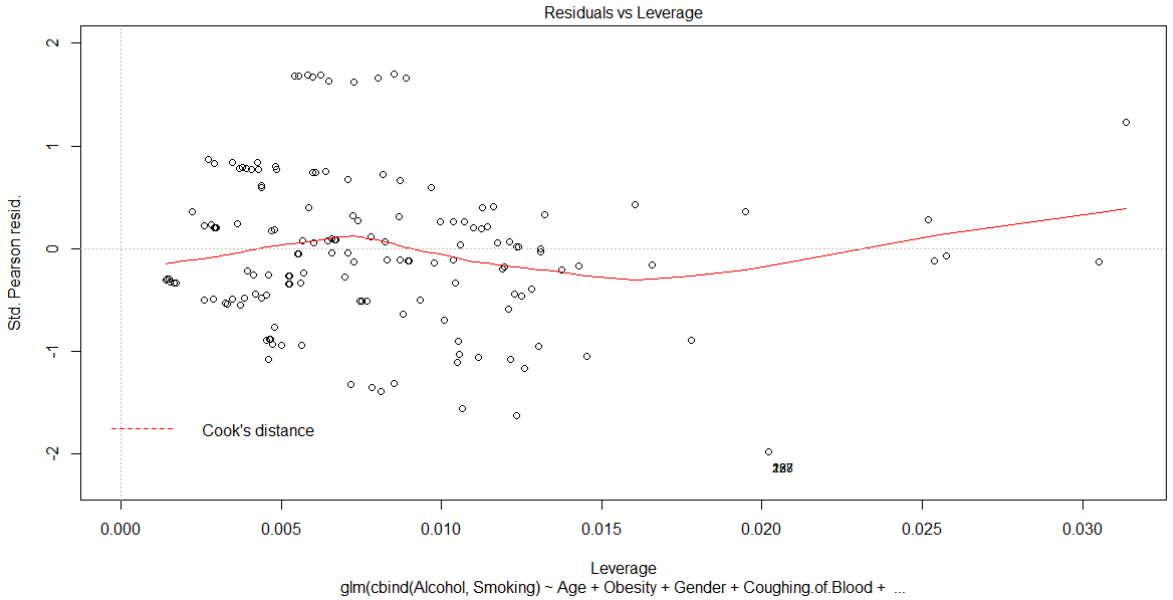
Null deviance: 773.77 on 999 degrees of freedom
Residual deviance: 587.86 on 992 degrees of freedom
AIC: 2999.2
Number of Fisher Scoring iterations: 4
I used the function cbind() function to create a matrix by binding the column vectors containing the use of alcohol and smoking.

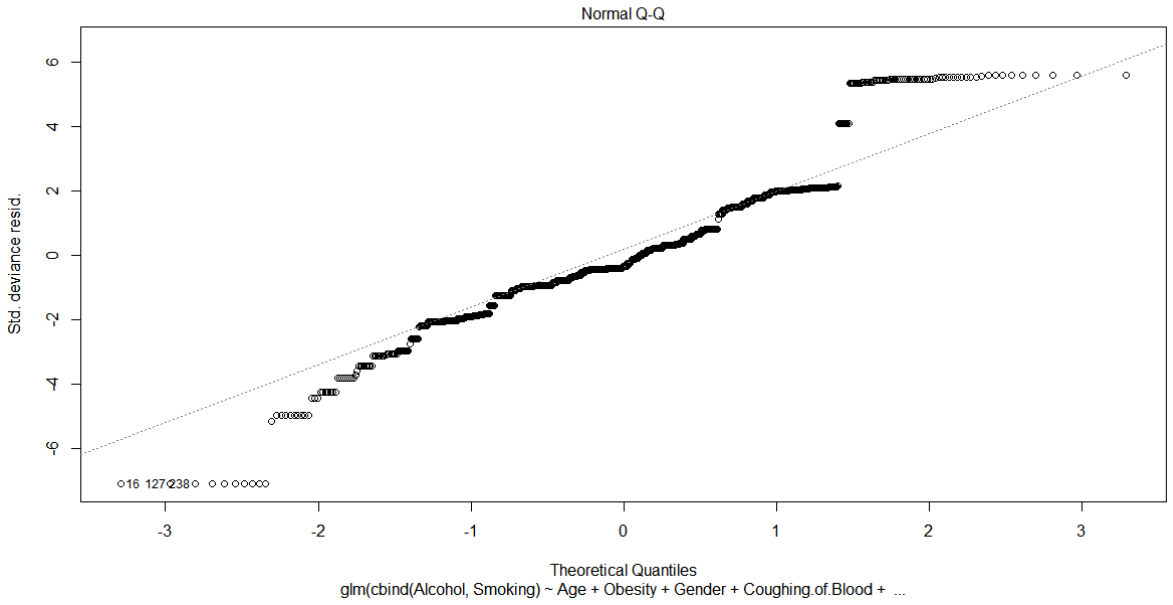
(Intercept) Age Obesity Gender
-0.7333487065 0.0008194979 0.0368608855 0.0380460432
Coughing.of.Blood Dry.Cough Levellow LevelMedium
0.0405177768 0.0507479858 -0.0104919574 0.6688713164

VI. CONCLUSION:

It is clear from the output that there is a significant relationship between chest pain of smoker and change of lung cancer since there is extremely high t-value of 17.483 and a p>|t| of 0%- which essentially means that this relationship has a near-zero change of being due to statistics variation or change.

Visualizing and output:
Having the regression summary output is important for checking the accuracy of the regression model and data to be used for estimation and prediction, but visualizing the regression is an important step to taken to communicate the results of the regression in a more digestive format.





REFERENCE:

[1]. http://cancerindia.org.in/lung-cancer/
[2]. http://www.newindianexpress.com/nation/2018/mar/11/cancer-threat-looming-large-over-north-east-pan-ne-rally-to-sensitize-masses-1785471.html
[3]. www.guwahatiplus.com/daily-news/assam-contributes-to-nearly-30-000-cancer-patients-every-year
[4]. https://timesofindia.indiatimes.com/city/guwahati/Assam-records-highest-number-of-cancer-cases-in-northeast/articleshow/18343751.cms

Prasenjit Nath." Diagnosis System for Lung Cancer using R Programming." IOSR Journal of Engineering (IOSRJEN), vol. 09, no. 12, 2019, pp. 52-55.