JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SOME DATA SCIENCE TECHNIQUES ON SURVIVAL LUNG CANCER

[1]A.Vani [2]T.Sukeerthi
[1]Academic Consultant [2]Academic Consultant
[1]Department of Statistics, Sri Padmavati Mahila Visvavidyalayam, Tirupati,
Andhra Pradesh,India

Abstract: Lung cancer is one of the leading causes of cancer-related mortality worldwide,necessitating the development of effective diagnostic and predictive models. This study utilizesa dataset from Kaggle, comprising 309 instances described by 16 attributes, to explore the application of Decision Trees and Binary Logistic Regression in lung cancer research. The objectives include building a Decision Tree to identify risk percentages, determining the most affected age group, assessing classification accuracy in Binary Logistic Regression, and identifying the most significant variables in the advanced stages of lung cancer. Decision Trees were employed to classify risk factors and predict disease progression, while Binary Logistic Regression was used to model the relationship between the dichotomous dependent variable (lung cancer presence) and independent variables. Analysis using SPSS software demonstrated that the risk for developing lung cancer is approximately 12.6%, with Allergy, Swallowing Difficulty, and Yellow Fingers identified as key predictors. Binary Logistic Regression further highlighted significant variables such as Chronic Disease, Fatigue, Coughing, and Smoking, with a prediction accuracy of 91.7%.These findings underscore the utility of Decision Trees and Binary Logistic Regression in lung cancer research, providing valuable insights for risk assessment and patient management. The study concludes that Yellow Fingers, in conjunction with other symptoms, is a significant predictor in the advanced stages of lung cancer.

Keywords: Lung cancer, Predictive Models, Decision Trees, Binary Logistic Regression, Risk.

I.INTRODUCTION:

Lung cancer represents a critical public health challenge globally, marked by its historical evolution from a rare disease to a predominant cause of cancer-related deaths. The rise in lung cancer incidence paralleled the widespread adoption of cigarette smoking in the early 20th century, with notable historical milestones underscoring its impact. Scientific consensus in the1950s definitively linked smoking to lung cancer, prompting significant public health initiatives and debates with tobacco companies. Concurrent advancements in surgical techniques, such as the pioneering pneumonectomy in 1933 and subsequent refinements, marked pivotal strides in treatment approaches, which evolved to encompass less invasive procedures by the mid-20th century.

The pathogenesis of lung cancer involves genetic mutations in lung cells, primarily triggered by cigarette smoke and other environmental carcinogens. Symptoms typically manifest in advanced stages, emphasizing the critical need for early detection through imaging and biopsy. Treatment strategies are multifaceted, tailored to cancer type, stage, and patient health. Surgical interventions like lobectomy and pneumonectomy, alongside chemotherapy, radiotherapy, and emerging immunotherapy, constitute primary modalities aimed at tumor eradication and symptom management. Despite ongoing challenges, preventive efforts center on smoking cessation and minimizing exposure to carcinogenic substances, underscoring the complex interplay between risk factors, diagnostic advancements, and therapeutic innovations in combating this pervasive disease.

II. REVIEW OF LITERATURE:

"Lung Cancer Death Rates Continue to Decline" (Journal of the National Cancer Institute, December 2023): This study examines the trends in lung cancer death rates in the U.S. from 1991 to 2018. The researchers found a significant decline in both smoking- attributable and smoking-unrelated lung cancer death rates. This decline is attributed to improved treatments, decreased smoking rates, and reduced exposure to secondhand smoke. The study highlights the need for ongoing public health efforts to reduce smoking prevalence further and address disparities in smoking-related cancer deaths across different communities.

"Lung Cancer Immunotherapy: Progress, Pitfalls, and Promises" (Molecular Cancer, 2023): This article provides an in-depth review of the progress in lung cancer immunotherapy, focusing on the genetic underpinnings and the unique mutational landscape of lung cancer. It discusses how specific oncogenic mutations and the immune system's role in targeting tumor cells have led to advancements in treatment. The review also addresses the challenges and future directions for immunotherapy, emphasizing the importance of personalized medicine based on genetic profiling of tumors.

"Lung Cancer Treatment Advances in 2023" (American Society of Clinical Oncology, ASCO): This review summarizes the latest treatment advancements for lung cancer,including the development of new targeted therapies and immunotherapy approaches. It highlights key clinical trials and the approval of new drugs that have shown promise in improving survival rates and quality of life for lung cancer patients. The article underscores the significance of ongoing research in identifying biomarkers that can predict response to thesetreatments.

III. METHODS AND MATERIALS:

3.1.Data Collection:

The data is carried out from Kaggle-data sets. https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection. The data set includes 309 instances. The instances are described by 16 attributes.

Objectives:

- To Build Decision Tree and to identify Risk percentage
- To Find which age group is most affected by Lung cancer.
- To Find the correctly classified cases in binary logistic regression.
- To Identify which variable is the most effected in fourth stage.

3.2. About The SPSS Software:

SPSS is powerful statistical software, which was earlier known for its applications in social sciences only. It is a comprehensive integrated system for flexible statistical data analysis and data management solution. SPSS is a computer program used for survey authoring and deployment, data mining, text analytics and collaborationand deployment. It has all major analytical tools for handling a large volume of data aswell as complicated multivariate analyses. The current version in use are IBM SPSS 16.0, 20.0.

3.3. Statistical Tools

We have used both Decision Trees and Binary Logistic Regression on this data Decision Trees are a non-parametric supervised learning method used for classification and regression. In the context of lung

cancer research, Decision Trees can effectively identify andclassify risk factors, predict disease progression, and assist in decision-making processes regarding patient treatment plans.

Binary Logistic Regression is a parametric statistical method used to model the relationship between a dichotomous dependent variable and one or more independent variables. In lung cancer research, this method is valuable for predicting the probability ofdisease occurrence and outcomes based on risk factors.

IV. RESULTS AND DISCUSSIONS:

Import the data from Excel to SPSS

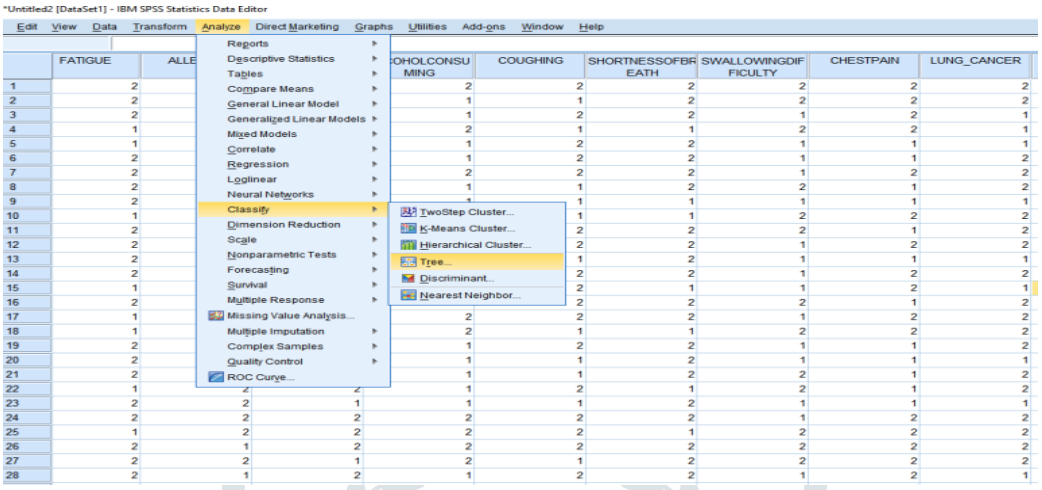


fig1: displays the analyze dialogue box

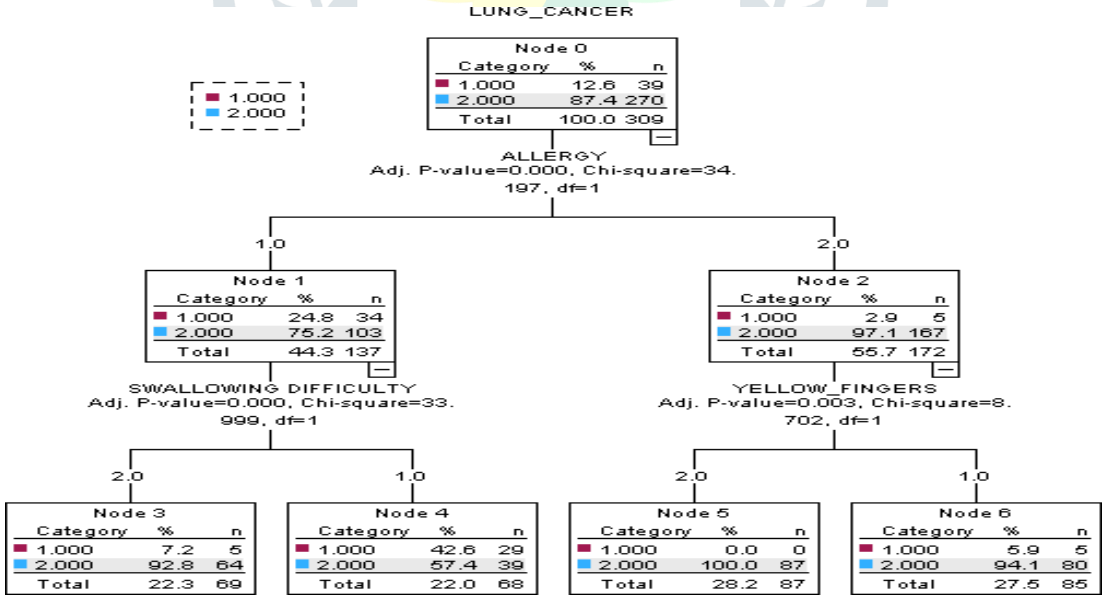


fig2: Decision Tree

Table 1: tree table

Node	1 = NO		2 = YES		Total		Primary Independent Variable				
	N	Percent	N	Precent	N	Percent	Variable	sig. a	Chi-sqaure	Df	Split values
0	39	12.6%	270	87.4%	309	100.0%	-	-	-	-	-
1	34	24.8%	103	75.2%	137	44.3%	Allergy	<.001	34.197	1	1.0
2	5	2.9%	167	97.1%	172	55.7%	Allergy	<.001	34.197	1	2.0
3	5	7.2%	64	92.8%	69	22.3%	Swallowing difficulty	<.001	33.999	1	2.0
4	29	42.6%	39	57.4%	68	22.0%	Swallowing difficulty	<.001	33.999	1	1.0
5	0	0.0%	87	100%	87	28.2%	Yellow fingers	.003	8.702	1	2.0
6	5	5.9%	80	94.1%	85	27.5%	Yellow fingers	.003	8.702	1	1.0

Using the CHAID method to analyze the predictors of lung cancer, the decision tree revealed that the overall risk for developing lung cancer in the dataset is 12.6%. Beginning with the rootnode displaying the distribution of lung cancer cases, the model splits the data based on the predictor with the strongest association, which in this case is "Allergy." This variable bifurcatesinto two child nodes: one where individuals do not have allergies (No), and another where theydo (Yes). Among those without allergies, the model predicts lung cancer in approximately 50.5% of cases across 309 individuals. Further down the tree, "Swallowing Difficulty" emerges as a significant predictor. For those without allergies and experiencing swallowing difficulties, the model predicts lung cancer in 44.3% of 137 cases, compared to 24.8% without swallowing difficulties. Similarly, "Yellow Fingers" is highlighted as crucial. Among those with allergies and yellow fingers, the model predicts lung cancer in 55.5% of 172 cases, contrasting with 2.9% for those without yellow fingers. Moving deeper, among individuals with swallowing difficulties, the model predicts lung cancer in 22.3% of 69 cases, versus 22% without such difficulties. Finally, among those without allergies and yellow fingers, the model predicts lung cancer in 28.2% of 85 cases, compared to 27.5% without yellow fingers, marking this as a terminal node in the decision tree analysis. This approach underscores the utility of CHAID inidentifying critical predictors of lung cancer risk based on specific patient characteristics and symptoms.

Table: 2 risk factor

RISK	
Estimate	Std. Error
.126	.019

Growing Method: CHAID, Dependent Variable: LUNG CANCERRisk for getting lung cancer is approximately 12.6%

Table: 3 classification

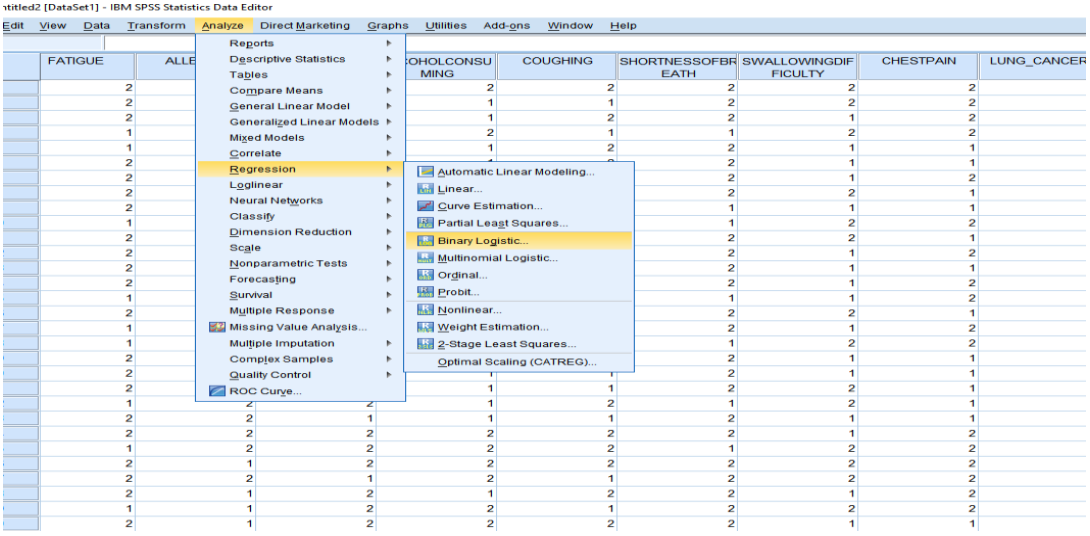
Classification			
Observed	**Predicted**		
	1=NO	**2=YES**	**Percent correct**
1=NO	0	39	0.0%
2=YES	0	270	100.0%
Overall percentage	0.0%	100.0%	87.4%

From the above classification table, we conclude that the overall 87.4% are correctlyclassified cases.

Binary Logistic Regression

Logistic Regression measures the relationship between the categorical target variable and one or more independent variables. It is useful for situations in which the outcome for a target Variable can have only two possible types (in other words, it is Binary).

Binary Logistic Regression classification makes use of one or more predictor variables that may be either continuous or categorical to predict the target variable classes. This technique helps to identify important factors (xi) impacting the target variable (Y) and also the nature ofthe relationship between each of



these factors and the dependent variable.

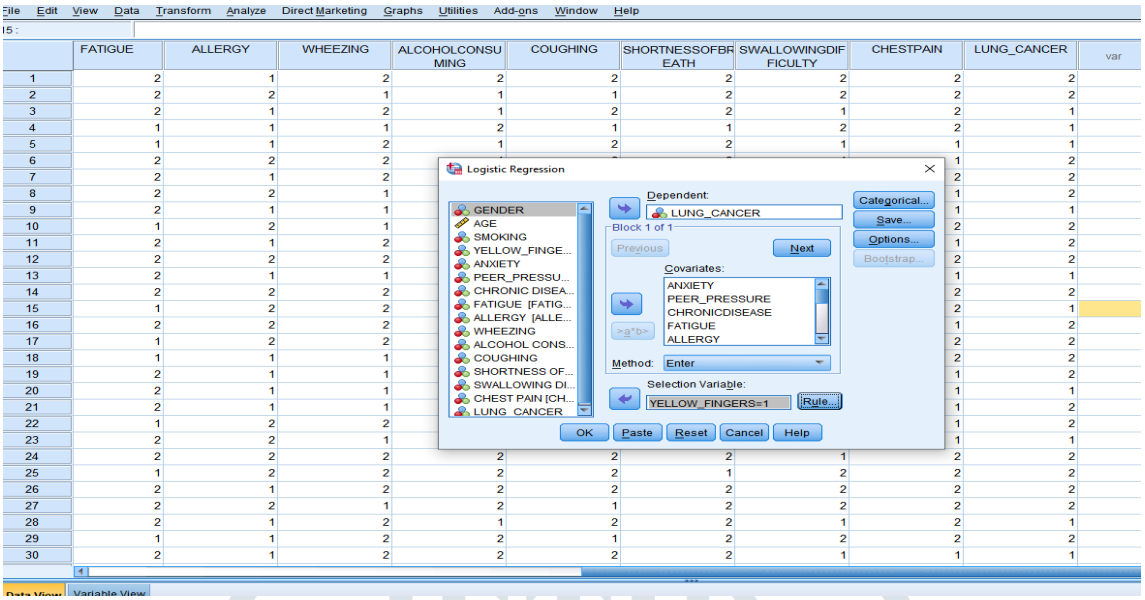


Table: 4 classification table

Classification Table								
			Predicted					
			Selected Cases			Unselected Cases		
			LUNG CANCER		Percentage Correct	LUNG CANCER		Percentage Correct
	Observed		1=NO	2=YES		1=NO	2=YES	
Step0	LUNG CANCER	1=NO	0	26	.0	0	13	.0
		2=YES	0	107	100.0	0	163	100.0
	Overall Percentage				80.5			92.6

Table: 5 variables equation

Variables in the Equation							
		B	**S.E.**	**Wald**	**Df**	**Sig.**	**Exp(B)**
Step 1	**AGE**	.032	.039	.665	1	.415	1.032
	ANXIETY	-2.112	1.281	2.716	1	.099	.121
	PEER- PRESSURE	2.157	1.192	3.277	1	.070	8.644
	CHRONICDISEASE	4.748	1.757	7.304	1	.007	115.330
	FATIGUE	4.087	1.512	7.308	1	.007	59.541
	ALLERGY	.123	1.072	.013	1	.908	1.131
	WHEEZING	.268	1.133	.056	1	.813	1.307
	ALCOHOL CONSUMING	2.452	1.266	3.754	1	.053	11.614
	COUGHING	3.284	1.397	5.529	1	.019	26.684
	SHORTNESSOF BREATH	-1.694	1.189	2.030	1	.154	.184
	SWALLOWING DIFFICULTY	5.970	2.467	5.855	1	.016	391.352
	CHEST PAIN	.602	1.060	.323	1	.570	1.826
	GENDER	-1.101	1.232	.799	1	.371	.332
	SMOKING	3.231	1.389	5.412	1	.020	25.300
	Constant	-32.68	9.528	11.053	1	<.001	.000

WALD STATISTIC FORMULAE: Wald = B/SEB

Table: 6 classification table

Classification Table								
			Predicted					
	Observed		Selected cases			Unselected cases		
			LUNG CANCER		Percentage Correct	LUNG CANCER		Percentage Correct
			1=NO	2=YES		1=NO	2=YES	
Step1	LUNG CANCER	1=NO	19	7	73.1	9	4	69.2
		2=YES	4	103	96.3	17	146	89.6
	Overall percentage				91.7			88.1

From the above classification table, it is evident that 91.7% of cases were correctly classified. In the context of the fourth stage, Yellow Fingers emerged as one of the most impactful variables. When predicting Yellow Fingers alongside other independent variables, significant associations were found with chronic disease (0.007), Fatigue (0.007), Coughing (0.019), Swallowing Difficulty (0.016) and Smoking (0.020). These findings underscore the critical role of Yellow Fingers as a predictor in advanced stages of the disease.

V. CONCLUSION:

Based on the analysis using Decision Trees and Binary Logistic Regression, several key findings regarding lung cancer risk and predictors have been identified. The Decision Tree analysis revealed that the overall risk for developing lung cancer in the dataset is estimated at 12.6%. Starting with the root node, which represents the initial distribution of lung cancer cases, the tree splits the data based on statistical significance, highlighting "Allergy" as the most influential predictor related to lung cancer. The model demonstrated strong performance, correctly classifying 87.4% of cases according to the classification table. Interestingly, the study found no significant association between age group and lung cancer presence. In contrast, Binary Logistic Regression yielded a higher accuracy rate of 91.7%, underscoring its effectiveness in predicting lung cancer based on various independent variables. Notably, in advanced stages of lung cancer (Stage 4), "Yellow Fingers" emerged as the most impactful predictor, significantly associated with other symptoms such as chronic disease, fatigue, coughing, swallowing difficulty, and smoking. This underscores the importance of Yellow Fingers as a diagnostic marker in identifying the advanced stages of lung cancer. Overall, the study emphasizes the utility of these analytical methods in understanding and predicting lung cancer risk and progression, particularly highlighting Yellow Fingers as a critical indicator in advanced disease stages.

VI. REFERENCES:

1. Adler, I. (1912). Primary Malignant Growths of the Lungs and Bronchi.
2. Ochsner, A. (1919). Personal Recollections of Lung Cancer.
3. British Doctor Study. (1954). Report on Smoking and Lung Cancer.
4. Wynder, E. L., & Graham, E. A. (1950). Tobacco Smoking as a Possible Etiologic Factor in Bronchiogenic Carcinoma.
5. National Cancer Institute. (2020). Lung Cancer Screening.
6. American Cancer Society. (2021). Tests for Lung Cancer.
7. McKenna, R. J., et al. (1994). VATS Lobectomy: Experience and Technique.
8. NSCLC Treatment Guidelines. (2021). American Society of Clinical Oncology.
9. SCLC Treatment Guidelines. (2021). National Comprehensive Cancer Network.
10. Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The Biology and Managementof Lung Cancer.
11. Rizvi, N. A., et al. (2015). Cancer Immunology and Immunotherapy: Lung Cancer.
12. Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in Relation to Smoking: 50 Years' Observations on Male British Doctors.
13. Jha, P., & Peto, R. (2014). Global Effects of Smoking, of Quitting, and of TaxingTobacco.
14. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification andRegression Trees. Chapman & Hall/CRC.
15. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied LogisticRegression. John Wiley & Sons.
16. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
17. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of StatisticalLearning. Springer.