Proceedings of Science and Mathematics
Faculty of Science,
Universiti Teknologi Malaysia

Vol. 16, 2023, page 86 – 97

Prediction of Lung Cancer Using Logistics Regression

Ea Li Qun, Haliza Abd Rahman*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: halizarahman@utm.my

Abstract

This research aims to analyze the presence of lung cancer in relation to the symptoms related to the disease. According to the World Health Organization (WHO), lung cancer is the third most common cancer and most worrying since it has the lowest survival rate among cancer, and 90 percent of cases go undetected until later stages. The findings of this study may help people that live in this endemic to better understand the disease and the factors that contribute to the worsening of the illness. The data used is a survey form collected from the public (*https://data.world/sta427ceyin/survey-lung-cancer*) which includes a total of 309 respondents. Analysis was done by using IBM SPSS Statistics Version 22 and Microsoft Office Excel 2019. Since the response variable for this study is of a binary nature which is the presence or absence of lung cancer, the logistics regression model was applied. Possible factors affecting the presence of lung cancer are gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain. The result shows that factors such as smoking, peer pressure, chronic disease, fatigue, allergy, coughing, and swallowing difficulty contribute significantly to the variation in presence of the lung cancer.

Keywords: Lung Cancer, Logistics Regression, Prediction Model, Disease

1. Introduction

The purpose of this study will produce a statistical profiling of lung cancer data and to apply the logistics regression to the medical data by predicting the presence of lung cancer and finding significant factors contributing to lung cancer.

Background of the study is cancer defined as a disease that develops when body tissue cells grow abnormally and transform into cancer cells, which are involved in tumor growth and infect the tissues around it. A tumor is a condition in which the body's cells grow abnormally. The genes that perform to influence the body's growth, development, or improvement are found in every cell that makes up the tissues in the human body. Eventually, the living cells will gather and combine with the newly formed cells. The cell will subsequently develop into a mass, which is known as a tumor [1]. Lung cancer is the most prevalent malignancy and the leading cause of cancer deaths globally in recent years. An estimated 1.8 million new cases, or 12.9% of all new cancer diagnoses, were recorded in 2012. The Global Hardship of Disease research in 2020 found that lung cancer had a significant global impact on healthcare costs and burdens. Its five-year survival rate (17.8%) was substantially lower than that of other malignant tumors. Due to the high fatality rate, the geographic patterns of death closely resemble those of incidence, and it is still a significant issue for public health [2]. The purpose of the use of logistics regression is wide, for example, it can be applied in the medical field. Developing lung cancer prediction models is essential. Logistic regression is currently the method used to create a predictive model. Hence, using machine learning approaches, we have found a method for identifying cancer in its early stages. In this work, the datasets will be classified using logistic regression. Since logistic regression is a generalized version of linear regression, we apply it to our study. It is worn mostly for processing dependent binary or multi-class variables. Since the response variable is discrete, linear regression cannot directly model it. Therefore, instead of forecasting the exact estimate of the occurrence, it forecast the odds of its occurring for the purpose of developing a model.

The problem statement of this study is lung cancer is the growth of abnormal cells in the lungs. This abnormal growth can occur in either one or both of the lungs. The abnormal cells cause the development of unhealthy lung tissue, which causes the lungs to not function properly. The main function of the lungs is to provide oxygen to the body via blood. Some of the symptoms of lung cancer are a persistent cough, shortness of breath, recurrent pneumonia, change in sputum, and coughing up blood. The lung is the second-most commonly diagnosed cancer in both men and women and is the most common cause of cancer death. In addition, the influence of factors on lung cancer has received substantial attention in their efforts to build a stable society. In this study, data relates to independent variables which believe to contribute to lung cancer will be obtained and applied to logistics regression. Logistics regression analysis will then be applied to the data to forecast the potential susceptibility to lung cancer.

This study is important in helping to analyze the significant factors which contribute to the presence of lung cancer. The medical and mathematical sectors will greatly benefit from the results of this study. We can confirm which factors are important to the diagnosis of lung cancer by applying the logistics regression approach in mathematics with the aid of the applied method. Better education on the awareness of lung cancer may also encourage more individuals to become alert in this day and age. People typically research the hazards when they have been affected by lung cancer for the reasons mentioned above. People frequently misunderstand the severity of the stages of lung cancer in particular since there has not been much research done on this topic. Additionally, information from this work can lessen the stigma in a community where lung cancer illness is only spread by elderly individuals. This stereotype poses a risk that it will make society today oblivious to impending disease.

The scope of the study is to study the concept of logistics regression and its application in lung cancer will be discussed. The concept of logistics regression is a prelude to the application in this study and will be applied in verifying the factors of lung cancer. The factors that will be considered due to lung cancer are gender, age, declaration of smoking status, types of systems, alcohol assumption, types of comorbidities, and allergic disorder. Microsoft Excel and SPSS will be used to analyze the data that has been collected. For each regression test that is run, the association between the variables will be determined using SPSS.

2. Literature Review

2.1 Lung Cancer

Tumors are disorders in which the body's cells grow abnormally, resulting in lesions or, in most cases, lumps in the body. There have two types of lung cancer that is benign tumors and malignant tumors. Benign tumors are noncancerous and emerge in lipomas, fibroids, or adenomas. Malignant tumors are cancerous tumors that tend to be abnormal such as breast cancer, lung cancer, and colorectal cancer. This study will concentrate on lung cancer since it has become the biggest cause of cancer-related death worldwide, with mortality rates exceeding those of prostate cancer, breast cancer, and cervical cancer. According to the World Cancer Research Fund International report, the second most frequent cancer in the world is lung cancer and there are more than 2.2 million new cases of lung cancer appeared in 2020.

According to GLOBOCAN 2022 data, the most prevalent cancers among both men and women were lung (19.2%), breast (10.8%), stomach (8.6%), liver (6.9%), and colon (6.0%). Lung cancer

(19.2%), liver (10.5%), stomach (9.9%), esophageal (7.5%), and breast (6.0%) cancers are the leading causes of cancer-related death in both men and women. Lung cancer is the most common cancer-related cause of death in Malaysia. As compared to other cancer types, lung cancer patients in Malaysia have the lowest survival rates [3].

2.2 Aspects that influence the severity of lung cancer

A chronic cough, bloody sputum, chest pain, vocal changes, increased shortness of breath, and recurring pneumonia or bronchitis are just a few of the symptoms that may appear before the malignancy has grown [4]. In addition, there are additional signs and symptoms including a persistent cough, hemoptysis (coughing up blood), hoarseness, cachexia (wasting condition; a decrease of body weight, muscle mass, and fat composition due to cancer), wheezing, chest pain, and clubbing of the fingers [5].

Gender appears to be a significant independent risk factor for developing lung cancer, which is the most common cancer to affect American women [6]. Epidemiological research shows that after considering cigarette consumption, women are three times more likely than males to have lung cancer. Age has been identified as a prognostic factor in many malignancies that are aggressively treated. Lung cancer is a common disease among the elderly because of prolonged life expectancies and an elevated risk of malignancy with aging. As compared to younger individuals, older patients have a higher incidence rate of development of squamous cell carcinomas in their lungs. Additionally, elderly individuals were found to be more frequent in the disease of anemia and they were less resistant to chemotherapy due to the presence of comorbidities [7].

In the analysis, it shows that alcohol consumption, especially beer consumption, is highly correlated to an increased risk of lung cancer [8]. Heavy drinkers have a 15% higher probability of developing lung cancer than either non-drinkers or occasional drinkers. Small cell lung cancer risk rose in non-smokers when alcohol consumption exceeded 0 - 4.9 g/day [9]. The patients with advanced lung cancer frequently have swallowing disorder problems which are also known as dysphagia. Individuals with advanced lung cancer receiving palliative chemotherapy are prone to swallowing issues, which may have an effect on their quality of life [10].

2.3 Logistics Modelling of lung cancer

Lung cancer became a prime utilization example of logistic regression from the analysis of relevant data [11]. In order to detect lung malignancies early, logistic regression and back propagation (BP) neural networks are utilized as prediction models. The R programming language will be used in the procedure to read and analyze statistical data related to lung cancer in order to create logistic regression models for the exploration of risk factors and the prediction of pain probabilities.

A logistic regression-based lung cancer prediction system that was used to classify the dataset. The dataset was split into two groups for training and testing first. In order to improve the system's accuracy to predict lung cancer in patients, a variety of training datasets are used to train the system using logistic regression. For an accurate outcome, the system is next evaluated using a testing dataset. Therefore, the raining accuracy was obtained as 96%, and the testing accuracy was 84% [4].

The accuracy precision, recall, F1 score, and confusion matrix are used to compare the performance of the Support Vector Machine (SVM) and Logistic Regression (LR) algorithms in predicting the survival rate of lung cancer patients based on age, clinical lung cancer stage, gender, histology, and dead status [12]. In order to make the most accurate predictions possible on lung cancer survivability, data cleaning, feature selection, splitting, and classification approaches have been used in this work. This study demonstrates that in comparison to the support vector machine classifier, which provides an accuracy of 76.20%, the logistic regression classifier provides the highest accuracy of

77.40%. Additionally, the logistic regression classifier provides the highest level of classification accuracy compared to all other classifiers.

3. Methodology

3.1 Data and Source

The data used in this study was obtained from one of the internet databases called data world, which collects information from a website that is Kaggle which includes a system for the online prediction of lung cancer. The data used is the raw lung cancer data; there are 309 rows of data and 16 attributes involved (*https://data.world/sta427ceyin/survey-lung-cancer*).

3.2 Logistics Regression Analysis Model

Logistics regression is the appropriate regression analysis to conduct when the dependent variables is dichotomous (binary). The logistics regression is a predictive analysis. Logistics regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. The simple logistics regression model is given by:

$$\beta_0 + \beta_1 x_1 = \log \pi - \log(1-\pi) = \frac{\log \pi}{1-\pi} \tag{1}$$

where π is a binomial proportion and x is explanatory variable β_0 and β_1 are the parameter of the logistics regression model. The multiple logistics regression model can be written as below:

$$\frac{\log \pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i \tag{2}$$

3.3 Univariate and Multivariate Analysis

Each explanatory variable will be modelled independently to the responding variable, Y, in logistic regression. The relationship between each predictor and Y variable will be examined in this analysis without consideration of any other variables. It considers only one variable in the analysis. Multivariate analysis, as opposed to univariate analysis, examines the relationship between every predictor and the responding variable, Y. Its goal was to study the connection between all predictors and Y. It is study the correlation between all predictors with y. In this process, the univariate analysis will be carried out for each variable. The variable that was tested to be significant will be fitted into the multivariate logistics regression model. Then, enter, backward, and forward model is carried out. Based on the classification accuracy, enter method regression analysis is carried out. With this method, all variables are entered into the model at once and it is done until there are only significant variables left in the model.

3.4 Wald Test

Wald test is used to find the significance of a coefficient which applied to the large-sample distribution of the maximum likelihood estimator. The hypothesis can be tested by

$$H_o: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

concerning the significance of a single coefficient by calculating the ratio of the estimate to its standard error

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

3.5 Hosmer and Lemeshow Test

The Hosmer and Lemeshow Test tells how good the model is. At each step, this is a goodness-of-fit test of the null hypothesis that the model adequately fits the data. If the null is true, the statistics should have an approximately chi-square distribution with the displayed degrees of freedom. If the significance level measure is more than 0.05, then do not reject H_0 and conclude that the model does fit the data very well. From each group the observed and expected numbers of events are computed for each group. The test statistics is

$$\hat{C} = \sum_{k=1}^{g} \frac{(Ok - E_k)^2}{V_k}$$

4. Results and Discussion

4.1 Descriptive Analysis

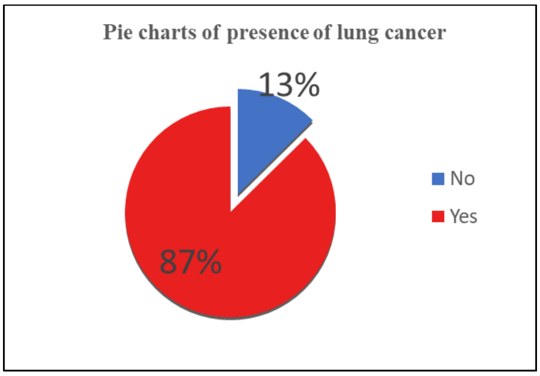


Figure 1 Pie chart of detecting the presence of lung cancer

Based on the data collection, 309 respondents were included in the study. The response variable is the presence of lung cancer while the predictor variable is the symptoms of lung cancer. Based on Figure 1 above, the presence of lung cancer is 87% while the absence of lung cancer was 13%.

4.2 Chi-square Test of Independence

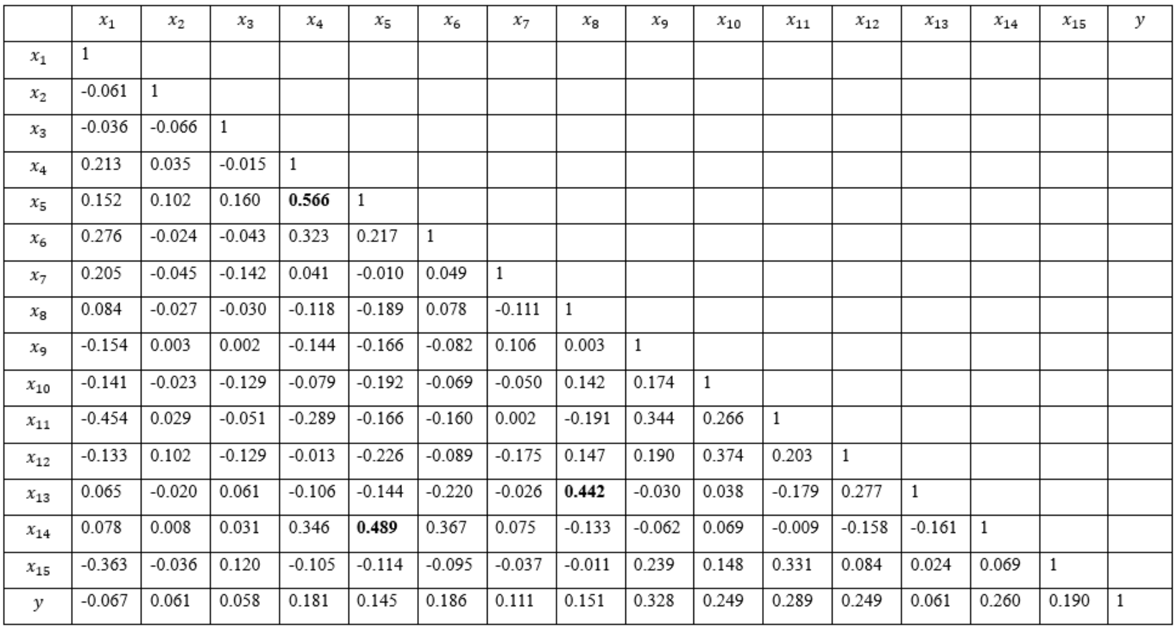
Table 1: Test of Independence

Number	Variables	Chi-square Test	df	p-value
1	Lung cancer and age	1.7530	2	0.4160
2	Lung cancer and gender	1.3980	1	0.2370
3	Lung cancer and smoking habits	1.0460	1	0.3060
4	Lung cancer and yellow fingers	10.1610	1	0.0010
5	Lung cancer and anxiety	6.4920	1	0.0110
6	Lung cancer and peer pressure	10.7350	1	0.0010
7	Lung cancer and chronic disease	3.8000	1	0.0510
8	Lung cancer and fatigue	7.0150	1	0.0080
9	Lung cancer and allergy	33.1960	1	<0.0001
10	Lung cancer and wheezing	19.2040	1	<0.0001
11	Lung cancer and alcohol consumption	25.7250	1	<0.0001
12	Lung cancer and coughing	19.0920	1	<0.0001

13	Lung cancer and shortness of breath	1.1400	1	0.2860
14	Lung cancer and swallowing difficulty	20.8450	1	<0.0001
15	Lung cancer and chest pain	11.2080	1	<0.0001

Based on Table 1, we could indicate that there was no association between the presence of lung cancer with age, gender, smoking habits, chronic disease, and shortness of breath since the p-value was greater than the significant value (α), 0.05. We conclude that there is a relationship between yellow fingers, anxiety, peer pressure, fatigue, allergy, wheezing, alcohol consumption, coughing, swallowing difficulty, and chest pain with the presence of lung cancer.

Table 2: Correlation Analysis between independent variables with dependent variables



	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	y
x_1	1															
x_2	-0.061	1														
x_3	-0.036	-0.066	1													
x_4	0.213	0.035	-0.015	1												
x_5	0.152	0.102	0.160	**0.566**	1											
x_6	0.276	-0.024	-0.043	0.323	0.217	1										
x_7	0.205	-0.045	-0.142	0.041	-0.010	0.049	1									
x_8	0.084	-0.027	-0.030	-0.118	-0.189	0.078	-0.111	1								
x_9	-0.154	0.003	0.002	-0.144	-0.166	-0.082	0.106	0.003	1							
x_{10}	-0.141	-0.023	-0.129	-0.079	-0.192	-0.069	-0.050	0.142	0.174	1						
x_{11}	-0.454	0.029	-0.051	-0.289	-0.166	-0.160	0.002	-0.191	0.344	0.266	1					
x_{12}	-0.133	0.102	-0.129	-0.013	-0.226	-0.089	-0.175	0.147	0.190	0.374	0.203	1				
x_{13}	0.065	-0.020	0.061	-0.106	-0.144	-0.220	-0.026	**0.442**	-0.030	0.038	-0.179	0.277	1			
x_{14}	0.078	0.008	0.031	0.346	**0.489**	0.367	0.075	-0.133	-0.062	0.069	-0.009	-0.158	-0.161	1		
x_{15}	-0.363	-0.036	0.120	-0.105	-0.114	-0.095	-0.037	-0.011	0.239	0.148	0.331	0.084	0.024	0.069	1	
y	-0.067	0.061	0.058	0.181	0.145	0.186	0.111	0.151	0.328	0.249	0.289	0.249	0.061	0.260	0.190	1

where
x_1 = Gender x_2 = Age x_3 = Smoking x_4 = Yellow fingers x_5 = Anxiety x_6 = Peer pressure
x_7 = Chronic disease x_8 = Fatigue x_9 = Allergy x_{10} = Wheezing x_{11} = Alcohol consuming
x_{12} = Coughing x_{13} = Shortness of breath x_{14} = Swallowing difficulty x_{15} = Chest pain
y = The presence of lung cancer

Table 2 above shows the correlation matrix for fifteen independent variables. From this table, there is indication of moderate positive correlation between anxiety and yellow fingers which is 0.566 and swallowing difficulty and anxiety which is 0.489. Another indication of moderate positive correlations is shortness of breath and breath. Other than that, there are indications of a weak correlation.

4.3 Multicollinearity

Table 3: Test of Multicollinearity

Variable	**Collinearity Statistics**	
	Tolerance	**VIF**
Age	0.953	1.049
Gender	0.662	1.511
Smoking	0.879	1.138
Yellow Fingers	0.557	1.795

Anxiety	0.493	2.030
Peer pressure	0.688	1.454
Chronic disease	0.850	1.177
Fatigue	0.686	1.457
Allergy	0.816	1.225
Wheezing	0.742	1.348
Alcohol Consumption	0.573	1.745
Coughing	0.652	1.533
Shortness of Breath	0.638	1.567
Swallowing Difficulty	0.620	1.614
Chest pain	0.770	1.298

Multicollinearity for all the independent variables was assessed by the tolerance and variation inflation factor (VIF). The result in Table 4.4 suggested weak collinearity between the independent variables as all the value of tolerance greater than 0.2. The VIF value for each of the independent variables is less than 5.0 which indicates that the variance of the estimated coefficients is not significantly inflated due to multicollinearity.

4.4 Univariate Analysis

Table 4: Univariate Analysis in presence of lung cancer

Variables	Wald statistics	df	p-value	Exp(B)	Adjusted OR (95% CI)
Age (1)	1.192	1	0.275	0.275	-1.359 (0.022, 2.947)
Age (2)	0.58	1	0.446	0.762	-0.272 (0.379, 1.533)
Gender	1.385	1	0.239	1.501	0..406 (0.763, 2.953)
Smoking	1.039	1	0.308	1.419	0.350 (0.724, 2.780)
Yellow fingers	9.484	1	0.002	3.047	1.115 (1.499, 6.191)
Anxiety	6.188	1	0.013	2.496	0.915 (1.214, 5.132)
Peer pressure	9.852	1	0.002	3.364	1.213 (1.577, 7.175)
Chronic disease	3.701	1	0.054	1.981	0.684 (0.987, 3.975)
Fatigue	6.714	1	0.010	2.456	0.899 (1.245, 4.847)
Allergy	23.503	1	<0.001	11.025	2.400 (4.178, 29.094)
Wheezing	16.509	1	<0.001	5.078	1.625 (2.319, 11.119)
Alcohol Consumption	20.497	1	<0.001	7.184	1.972 (3.059, 16.868)
Coughing	16.599	1	<0.001	4.852	1.579 (2.270, 10.374)
Shortness of breath	1.131	1	0.288	1.447	0.369 (0.733, 2.858)
Swallowing difficulty	16.23	1	<0.001	7.323	1.991 (2.780, 19.292)
Chest pain	10.358	1	0.001	3.273	1.186 (1.590,6.737)

At univariate level, it is found that ten variables are significant as the p-values less than 0.50 and the 95% confidence interval of the odds ratio did not contain the value of 1, namely yellow fingers, anxiety, peer pressure, fatigue, allergy, wheezing, alcohol consumption, coughing, swallowing difficulty, and chest pain. All significant variables from Table 4 are then included in the multivariate model where purposeful selection method are performed. At the multivariate level, a variable was considered significant if the p-values less than 0.05. The result of the best multivariate logistics model for the lung cancer patient is given in Table 5.

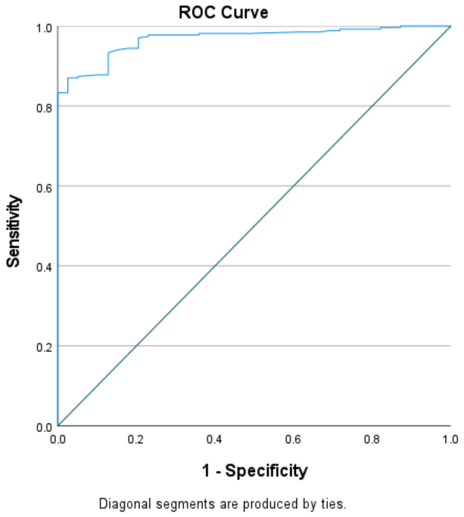
4.5 Multivariate Analysis

Table 5: Multivariate Analysis in presence of lung cancer

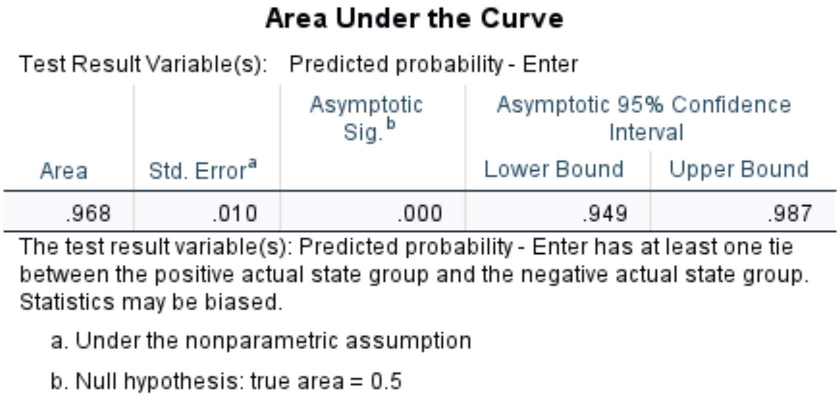
	Method 1: Enter	Method 2: Forward	Method 3: Backward
Omnibus Test- Chi-Square	144.724 (<0.001)	118.877 (<0.001)	137.669 (<0.001)
-2LL	89.575	115.422	96.631
Nagelkerke R Square	0.704	0.601	0.676
Classification Accuracy	95.1%	91.9%	94.5%
Hosmer and Lemeshow Test			
Chi-square	3.272 (0.916)	3.710 (0.882)	1.461 (0.993)
Constant	15.913	9.589	14.101
Gender (1)	-0.585 (0.420)	-	-
Age (1)	-2.822 (0.094)	-	-
Age (2)	0.168 (0.799)	-	-
Smoking (1)	-1.826 (0.012)	-	-1.454 (0.026)
Yellow Fingers (1)	-1.401 (0.062)	-1.784 (0.005)	-1.741 (0.006)
Anxiety (1)	-0.839 (0.308)	-	-
Peer Pressure (1)	-1.940 (0.005)	-1.569 (0.005)	-1.874 (0.003)
Chronic Disease (1)	-3.301 (<0.001)	-	-2.695 (<0.001)
Fatigue (1)	-3.294 (<0.001)	-2.271 (<0.001)	-2.870 (<0.001)
Allergy (1)	-1.892 (0.022)	-2.422 (<0.001)	-1.834 (0.011)
Wheezing (1)	-1.068 (0.204)	-	-
Alcohol Consuming (1)	-1.506 (0.067)	-1.888(0.003)	-1.751 (0.014)
Coughing (1)	-3.234 (0.003)	-1.119 (0.052)	-3.065 (<0.001)
Shortness of Breath (1)	0.711 (0.358)	-	-
Swallowing Difficulty (1)	-3.142 (0.006)	-2.005(<0.001)	-3.427 (<0.001)
Chest pain (1)	-0.461 (0.516)	-	-

Based on Table 5, the Logistics Regression Model using enter method is good for providing a prediction. In this case, the value of Chi-square for this model was 144.724 ($p-value < 0.05$) and indicate the model is good fit for the data. Besides, the value of -2 Log likelihood which is 89.575 smaller than the value of forward and backward method shows a good fit for data. Next, the values of Nagelkerke R^2 is 0.704 and it means 70.4% of variability accounted for by all of the predictors together. Based on Hosmer and Lemeshow test, the model adequately fits the data since the value is 3.272 and ($p-value > 0.05$). Accurate classification is 95.1% whereas the error rate is only 4.9%. Note that the overall correctly classified without any statistical predictions at all was 87.4%. It can be seen that it has increased the overall accuracy and therefore there is an improvement in goodness of fit by adding all of these predictors. Of these 309 cases, 39 respondents do not have lung cancer and 30 of these were correctly predicted since it has a fitted probability of less than 0.5. Similarly, 264 out of 270 observed have lung cancer since it has a fitted probability of more than 0.5. Based on the enter method, only smoking, peer pressure, chronic disease, fatigue, allergy, coughing and swallowing difficulty are significant at the 5% level of significance which means that there is statistical evidence that these variables contributes significantly to the presence of the lung cancer. Otherwise, gender, age, yellow fingers, anxiety, wheezing, alcohol consuming, shortness of breath and chest pain are not significant since there is no statistical evidence from the analysis.

4.6 Goodness of fit of the model



Diagonal segments are produced by ties.



Area Under the Curve

Test Result Variable(s): Predicted probability - Enter

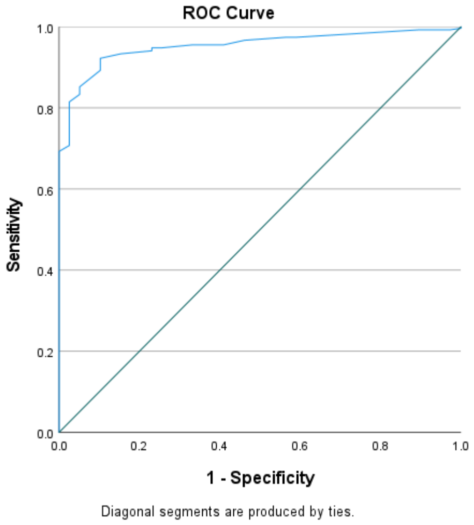
		Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
Area	Std. Error[a]		Lower Bound	Upper Bound
.968	.010	.000	.949	.987

The test result variable(s): Predicted probability - Enter has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

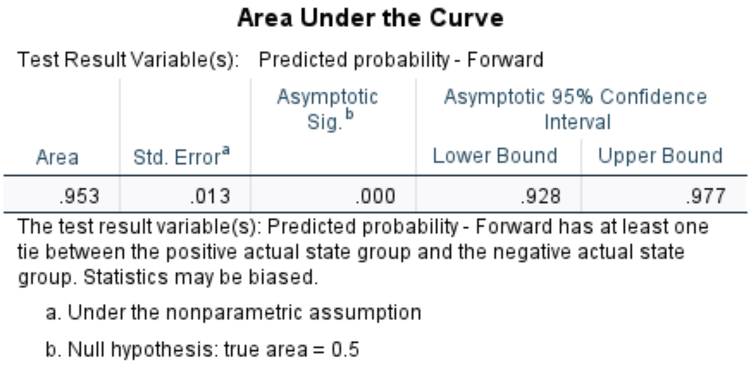
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Figure 2 ROC curve and AUC value for Enter method



Diagonal segments are produced by ties.



Area Under the Curve

Test Result Variable(s): Predicted probability - Forward

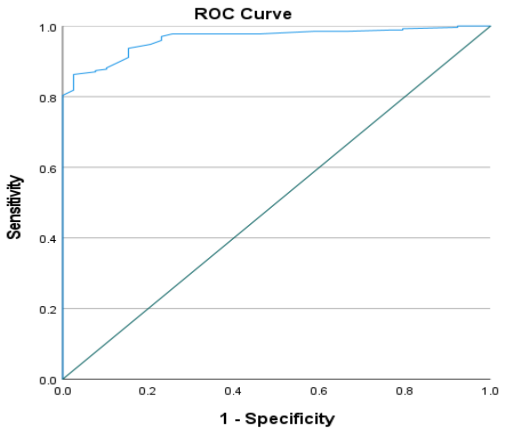
Area	Std. Error^a	Asymptotic Sig.^{b}	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.953	.013	.000	.928	.977

The test result variable(s): Predicted probability - Forward has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

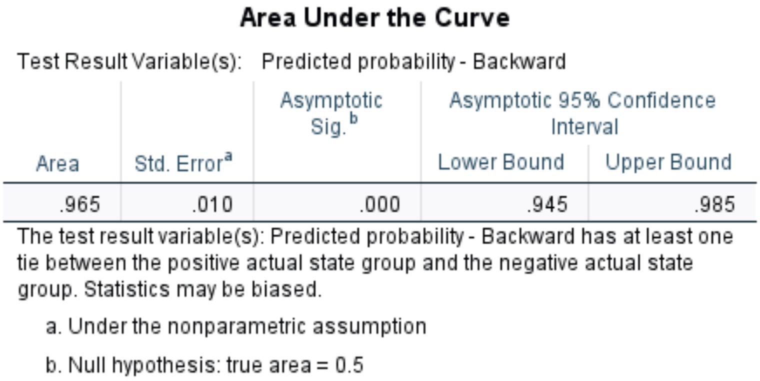
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Figure 3 ROC curve and AUC value for Forward method



Diagonal segments are produced by ties.



Area Under the Curve

Test Result Variable(s): Predicted probability - Backward

Area	Std. Error^a	Asymptotic Sig.^{b}	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.965	.010	.000	.945	.985

The test result variable(s): Predicted probability - Backward has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

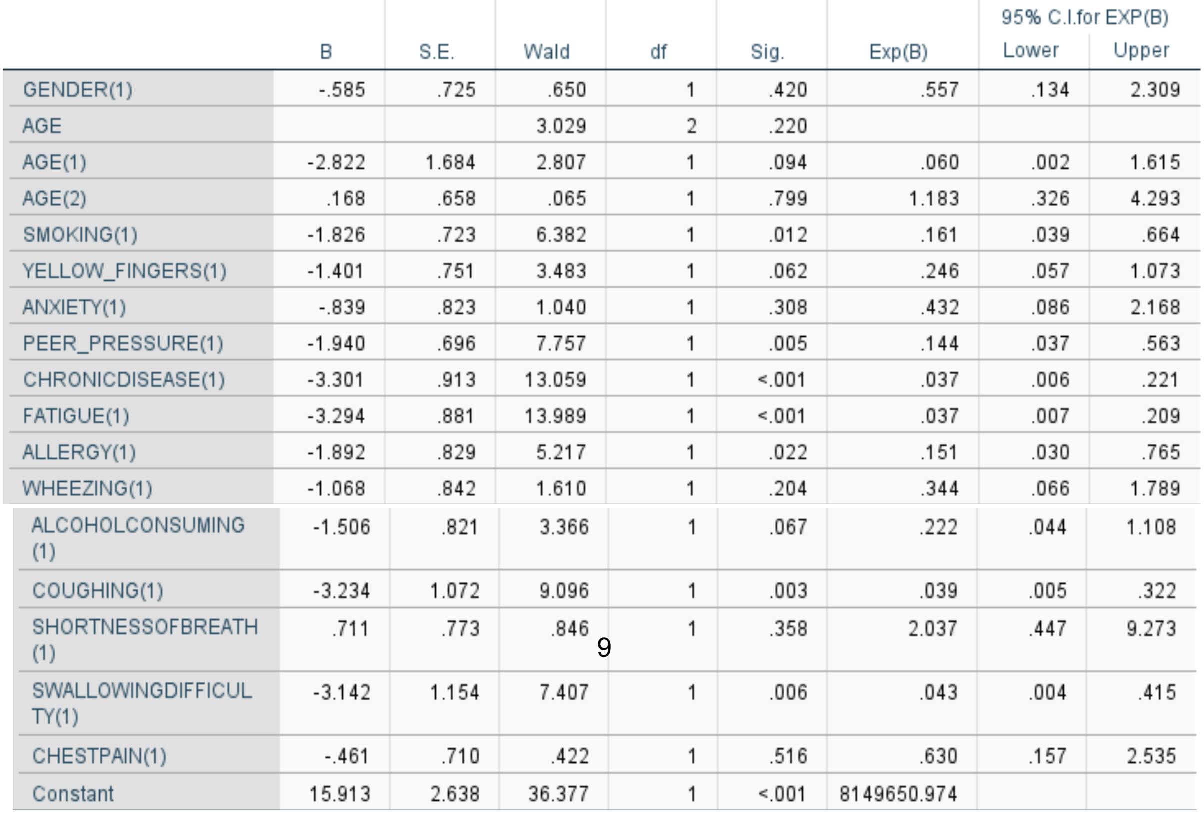
Figure 4 ROC curve and AUC value for Backward method

Based on Figure 2,3,4, the area under the Receiver Operating Characteristics (ROC) curve was 0.968 for the enter method, 0.953 for the forward method and 0.965 for the backward method respectively. It shows the probability that the result for the a randomly chosen positive case (presence of lung cancer) exceeded the result for a randomly chosen negative case (absence of lung cancer). Since the p-value is less than α = 0.05 for all three methods, then we can conclude all models are significant model for predicting the presence of lung cancer of the patients. The enter model has the highest AUC, which indicates that it has the highest area under the curve and is the best model at correctly classifying observations into categories.

4.7 Establish the final model

By taking into considerations of all statistical analysis and since the classification accuracy for enter method is the largest compared to the forward and backward method, therefore, model with the enter method is chosen as the bet estimated model for the perceptions on presence of lung cancer of the patients. Therefore, the estimated logistics model is

Table 6: Final Model by using Enter Logistics Regression



	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
GENDER(1)	-.585	.725	.650	1	.420	.557	.134	2.309
AGE			3.029	2	.220			
AGE(1)	-2.822	1.684	2.807	1	.094	.060	.002	1.615
AGE(2)	.168	.658	.065	1	.799	1.183	.326	4.293
SMOKING(1)	-1.826	.723	6.382	1	.012	.161	.039	.664
YELLOW_FINGERS(1)	-1.401	.751	3.483	1	.062	.246	.057	1.073
ANXIETY(1)	-.839	.823	1.040	1	.308	.432	.086	2.168
PEER_PRESSURE(1)	-1.940	.696	7.757	1	.005	.144	.037	.563
CHRONICDISEASE(1)	-3.301	.913	13.059	1	<.001	.037	.006	.221
FATIGUE(1)	-3.294	.881	13.989	1	<.001	.037	.007	.209
ALLERGY(1)	-1.892	.829	5.217	1	.022	.151	.030	.765
WHEEZING(1)	-1.068	.842	1.610	1	.204	.344	.066	1.789
ALCOHOLCONSUMING (1)	-1.506	.821	3.366	1	.067	.222	.044	1.108
COUGHING(1)	-3.234	1.072	9.096	1	.003	.039	.005	.322
SHORTNESSOFBREATH (1)	.711	.773	.846	1	.358	2.037	.447	9.273
SWALLOWINGDIFFICULTY(1)	-3.142	1.154	7.407	1	.006	.043	.004	.415
CHESTPAIN(1)	-.461	.710	.422	1	.516	.630	.157	2.535
Constant	15.913	2.638	36.377	1	<.001	8149650.974		

$$\text{Log}\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = 15.913 - 1.826 * \text{Smoking} - 1.940 * \text{Peer pressure} - 3.301 * \text{Chronic disease} - 3.294 * \text{Fatigue} - 1.892 * \text{Allergy} - 3.234 * \text{Coughing} - 3.142 * \text{Swallowing difficulty}$$

The interpretations of the coefficients of the significant variables that can be drawn based on Table 5 are each one unit decrease in smoking, the odds of the presence of lung cancer among the patients is increased by 0.161. Every one unit increase in smoking on the odds of being presence of lung cancer lies between 0.039 and 0.664 at 95% of the confidence interval. Each one unit decrease in peer pressure, the odds of the presence of lung cancer among the patients is increased by 0.144. Every one unit increase in peer pressure on the odds of being presence of lung cancer lies between 0.037 and 0.563 at 95% of confidence interval. Besides that, each one unit decrease in chronic disease, the odds of presence of lung cancer among the patients is increased by 0.037. Every one unit increase in chronic disease on the odds of being presence of lung cancer lies between 0.006 and 0.221 at 95% of confidence interval. The same interpretation was made on the variable fatigue, coughing and swallowing difficulty.

Conclusion

We can see more clearly comparison between the variable selection which are forward selection, backward elimination and enter method. Smoking, peer pressure, chronic disease, fatigue, allergy, coughing, and swallowing difficulty are more likely as the risk factors of lung cancer. The advantage of logistic regression is its usefulness for situation in which there is two or more outcomes based on set of predictor variables. This method is the most popular analysis that has been used for modelling dichotomous dependent variables. Since the response variable for this study is a binary nature which is the presence of lung cancer, the logistics regression was used to develop the model.

References

[1] Priscilia Lovita Paelongan, & Palupi, I. (2022). Lung Cancer Prediction Model using Logistic Linear Regression with Imbalanced Dataset. Indonesia Journal on Computing (Indo-JC), 7(2), 1–14. https://doi.org/10.34818/INDOJC.2022.7.2.616

[2] Wong, M. C. S., Lao, X. Q., Ho, K.-F., Goggins, W. B., & Tse, S. L. A. (2017). Incidence and mortality of lung cancer: global trends and association with socioeconomic status. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-14513-7

[3] Pathmanathan Rajadurai, Soon Hin How, Liam, C.-K., & Lye Mun Tho. (2020, March). Lung Cancer in Malaysia. ResearchGate; Lippincott, Williams & Wilkins.

[4] Raghavendra, Patil G E., Sinchana, C G., Tejashwini, P., et al. 2020. Lung Cancer Prediction System Using Logistic Regression Approach. International Research Journal of Modernization in Engineering Technology and Science, 656-660.

[5] Dewi, A., Thabrany, H., Satrya, A., Chairunnisa, G., Rifqi, P., Fattah, A., Novitasari, D., Paru, K., Paling, K., Di Indonesia, M., Saja, A., Telah, Y., Atasi, K., Apa, D., Kita, Y., & Lakukan, B. (2021). Lv, Y. and Gao, J. (2011) ‘Condition prediction of chemical complex systems based on Multifractal and Mahalanobis-Taguchi system’, in *ICQR2MSE 2011 - Proceedings of 2011 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, pp. 536–539.

[6] Gasperino, J. (2011). Gender is a risk factor for lung cancer. Medical Hypotheses, 76(3), 328–331. https://doi.org/10.1016/j.mehy.2010.10.030.

[7] TAS, F., CIFTCI, R., KILIC, L., & KARABULUT, S. (2013). Age is a prognostic factor affecting survival in lung cancer patients. Oncology Letters, 6(5), 1507–1513. https://doi.org/10.3892/ol.2013.1566

[8] Brenner, D. R., Fehringer, G., Zhang, Z.-F., Lee, Y. T., Meyers, T. J., Matsuo, K., Ito, H., Paolo Vineis, Stücker, I., Paolo Boffetta, Brennan, P., Christiani, D. C., Diao, N., Hong, Y.-C., Maria Teresa Landi, Morgenstern, H., Schwartz, A. G., Rennert, G., Saliba, W., & McLaughlin, J. R. (2019). Alcohol consumption and lung cancer risk: A pooled analysis from the International Lung Cancer Consortium and the SYNERGY study. 58, 25–32. https://doi.org/10.1016/j.canep.2018.10.006.

[9] Bagnardi V, Rota M, Botteri E, et al. Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. Br J Cancer 2015;112(3):580–593.

[10] Brady GC, Roe JWG, O’Brien M, Boaz A, Shaw C. An investigation of the prevalence of swallowing difficulties and impact on quality of life in patients with advanced lung cancer. Support Care Cancer 2018, 26(2):515–9

[11] Shen, S., Fan, Z., & Guo, Q. (2017). Design and application of tumor prediction model based on statistical method. Computer Assisted Surgery, 22(sup1), 232-239.

[12] Animesh Hazra, Bera, N., & Mandal, A. (2017, September 15). Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms. ResearchGate; Foundation of Computer Science.