DOTE5110 (IT)

Statistical Analysis

Evaluation of Key Factors

Affecting Lung Cancer Prediction Probability

Group Members

Haipo Liu 1155218447

Xinger Lin 1155216531

Yuying Shi 1155215015

Qiang Xu 1155220556

Xuan Li 1155215838

0. Introduction

Lung cancer is the leading cause of cancer deaths (18%) globally, making research on lung cancer critical to public health. This study analyzed a Kaggle dataset[1] containing lifestyle habits and health-related characteristics to identify risk factors associated with lung cancer.

This analysis integrated multiple variables such as smoking habits, age, and health characteristics to reveal the key factors and complex interactions that influence lung cancer risk using regression analysis and machine learning.

1. Data Cleaning

a. Handling Duplicate Values: Duplicate rows were detected and removed, resulting in the elimination of 33 instances to ensure data integrity.
b. Handling Missing Values: A thorough check for missing values confirmed the absence of any 'NA' entries in the dataset.
c. Column Name Standardization: Column names were standardized by removing leading and trailing spaces and replacing spaces with underscores for consistency.
d. Data Type Conversion: Categorical variables were transformed into a more readable format, with conversions made to numeric values in the original dataset for subsequent regression analyses and machine learning applications.

2. Data Visualization

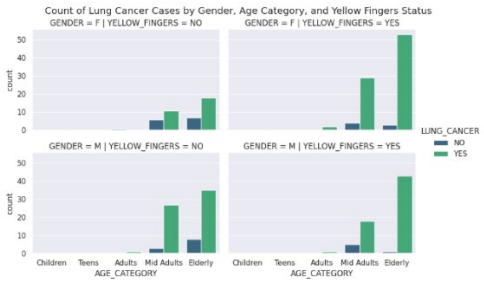
a. Overall Description: The dataset includes 276 samples with balanced genders (mean 0.49) and a mean age of 63. Smoking prevalence is 54%, while 86% have lung cancer diagnosis. The data show high dispersion, with std near 0.5, indicating good representation and accuracy.
b. Lung Cancer and Age/gender: Distribution graphs and histograms were created to analyze age/gender distribution in relation to lung cancer status (yes/no). And find Lung cancer is concentrated in individuals aged 55 to 70.
c. Cross-Analysis of Variables:
   Compare the distribution of cases under different variable scenarios (T/F) using customized age segments as x-axis and number of illnesses as y-axis

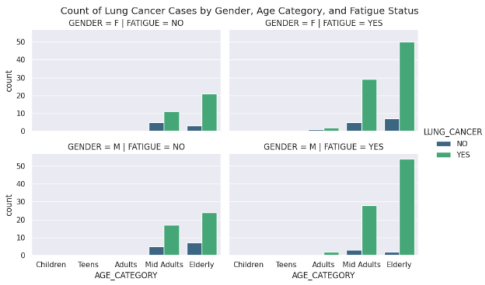
The name of the variable to be multiplied by gender				
Smoking	Yellow_Fingers	Anxiety	Peer_Pressure	Chronic_Disease
Fatigue	Allergy	Wheezing	Alcohol_Consuming	Coughing
Shortness_Of_Breath		Swallowing_Difficulty		Chest_Pain

Findings:

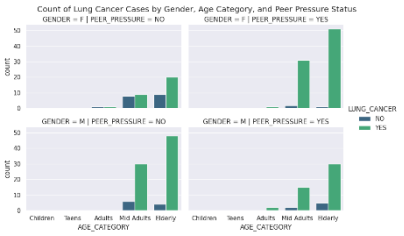
1. From a numerical standpoint, the variables [Yellow_Fingers, Fatigue, Coughing, Shortness_Of_Breath] have a significant impact on lung cancer (as shown below). The number of patients with these symptoms is significantly higher.

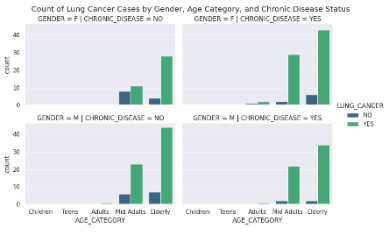
1 https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer, Collected from the online lung cancer prediction system .

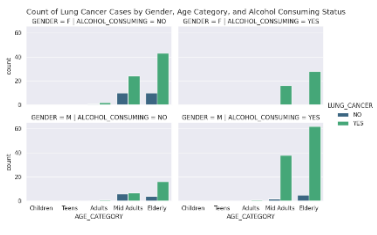




2.From a numerical perspective, the variables [Peer_Pressure, Chronic_Disease, Alcohol_Consuming] show gender differences in their impact, requiring further evaluation

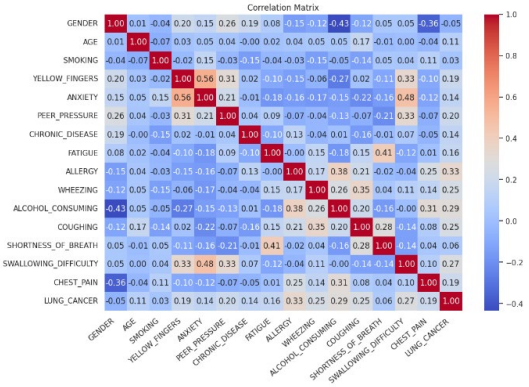






d. Correlation coefficient matrix

The correlation matrix identifies the top factors positively correlated with lung cancer: allergy (0.33), alcohol_consuming (0.29), swallowing_difficulty (0.27).etc. In contrast, gender (-0.05) and age (0.11) show minimal correlation, indicating that lifestyle and health status are more significant in lung cancer development.



3. Regression

a. Data preparation

Firstly, we select the most 5 highly correlated variables based on corr, ALLERGY, ALCOHOL_CONSUMING, SWALLOWING_DIFFICULTY, COUGHING, WHEEZING.

b. Logistic Regression

i. Logistic Regression Results

From this picture, we can know that all the p values are less than 0.05, so we consider all the five variables have a significant impact on determining the lung cancer probability. Also, according to the coef, we can see that all the 5 variables have a positive correlation

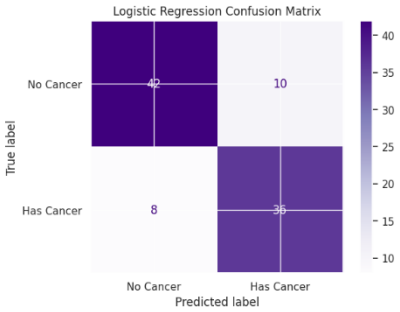
with lung cancer probability. Among these 5 variables, x1 and x5 have a coef of more than 1, so the impact is greater.

Logit Regression Results

Dep. Variable:	LUNG_CANCER	No. Observations:	380
Model:	Logit	Df Residuals:	374
Method:	MLE	Df Model:	5
Date:	Fri, 08 Nov 2024	Pseudo R-squ.:	0.4604
Time:	08:00:57	Log-Likelihood:	-142.08
converged:	True	LL-Null:	-263.31
Covariance Type:	nonrobust	LLR p-value:	2.271e-50

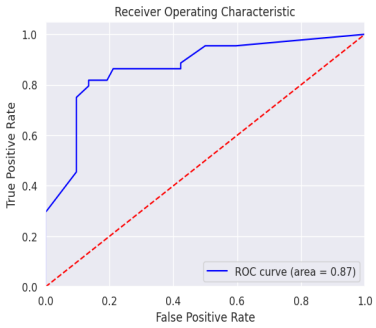
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4628	0.213	6.877	0.000	1.046	1.880
x1	1.1716	0.174	6.722	0.000	0.830	1.513
x2	0.4040	0.170	2.372	0.018	0.070	0.738
x3	0.5298	0.161	3.298	0.001	0.215	0.845
x4	0.6909	0.177	3.907	0.000	0.344	1.037
x5	1.2547	0.180	6.967	0.000	0.902	1.608

ii. Confusion Matrix



Here we use accuracy to evaluate the performance of the model. The accuracy is 0.81, indicating that our model correctly classified 81% of the samples.

iii. ROC curve:



As it can be seen from the figure, the curve is above the diagonal most of the time, indicating that the model has a good classification ability. And since the AUC value is 0.87, it indicates that the model performs well in distinguishing between positive samples.

c. LASSO Regression

i. LASSO Regression Result

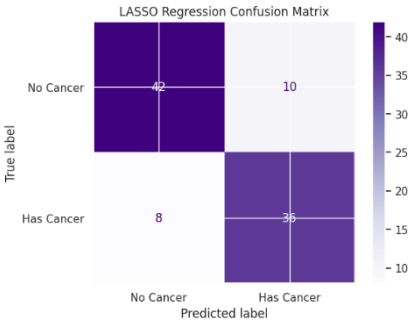
By using LASSO Regression model, we get the result
LASSO coef: [0.16939373 0.05984559 0.08466814 0.09930279 0.17717932],
which means all the 5 variances have a positive correlation with lung cancer probability, and the last one has the greatest impact. It implicates that WHEEZING impacts the lung cancer probability most. In this analysis, we observe that the two regressions agree, which is because the LASSO regression is actually an extended version of the Ordinary Least Squares (OLS)

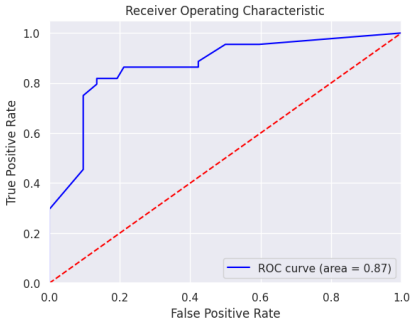
$$\min_{\beta}\left[-\frac{1}{N}\sum_{i=1}^{N}\left(y_i log(p_i)+(1-y_i)log(1-p_i)\right)+\lambda\sum_{j=1}^{p}|\beta_j|\right]$$

method. When the regularisation parameter lambda of LASSO is 0, the regression results are identical to those of OLS.

ii. Model evaluation

the Confusion Matrix and ROC curve are as followings





4. Machine Learning

a. Data Preparation

We use 80% of the data as a training set and 20% as a test set here.

b. SVM:

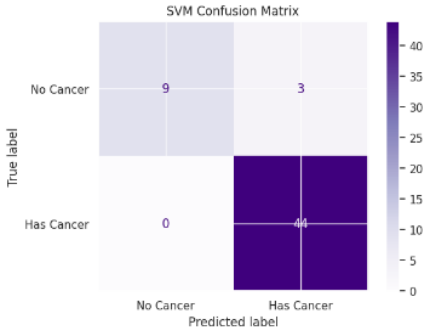
i. Confusion Matrix:

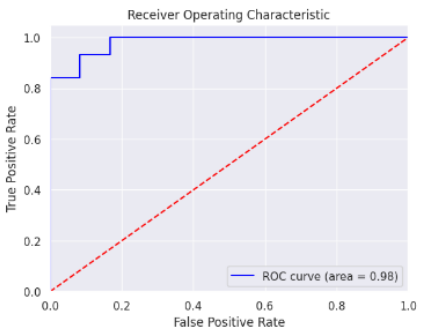
By training SVM model, in the test set we find that 44 patients who have cancer are correctly classified to “Has Cancer”, and 9 samples who have no cancer are also correctly classified to “No Cancer”. Also, 3 samples that the model incorrectly predicted 3 samples that were “No Cancer” as “Has Cancer”. This is called a false positive and indicates that the model has made some bad judgement in this area.

And our model has an accuracy of 0.95, which basically correctly predicts whether or not a patient has lung cancer.

ii. ROC curve:

Our SVM model showed excellent performance on the test set with an AUC value of 0.981, indicating a very high accuracy in distinguishing between “No Cancer” and “Has Cancer” patients.The curve has a stepped shape,which usually implies that the model has a high classification ability under some specific decision thresholds.





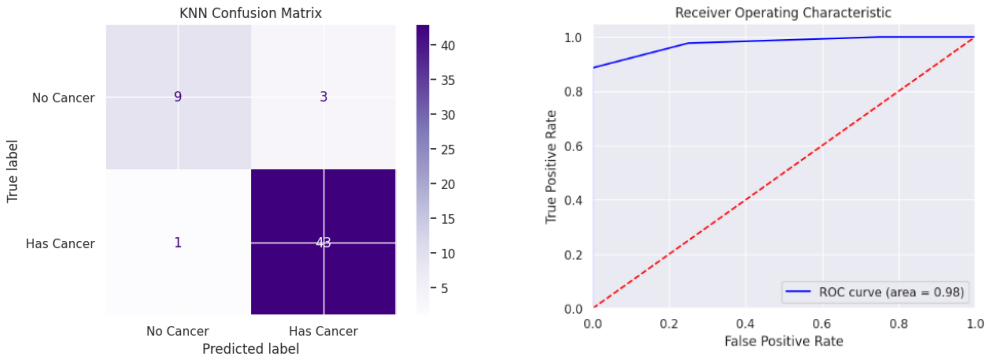
c. KNN:

i. Confusion Matrix:

In this model, there is 1 False Negative sample, which means the model incorrectly predicted a sample with “Has Cancer” as “No Cancer”.This means that in the real world, the model may have a missed diagnosis, which is a concern in medical applications. In all ,this model’s accuracy is 0.93, suggests that the model performs well on the problem of effectively differentiating between patients with and without cancer.

ii. ROC curve:

An AUC of 0.98 indicates that the performance of the model is excellent, recognizing both positive and negative classes almost perfectly.



5. Result and Conclusion

We fing the relation that the significant correlation between lifestyle factors (e.g., smoking, yellow fingers, alcohol consumption, and dysphagia) and lung cancer suggests that these factors may play an important role in the development of lung cancer.Sex and age had a smaller effect on lung cancer, possibly indicating that other lifestyle and health characteristics are more important in the development of lung cancer. And we consider allergy (0.33), alcohol_consuming (0.29), swallowing_difficulty (0.27),coughing and wheezing(0,25) as top 5 factors affecting lung cancer.

Except SVM and KNN ,we also train other machine learning models which contain Random Forest and GradientBoostingClassifier, we compare the performance of the 6 models and we draw a conclusion that the best method is Support Vector Machine and the best accuracy is 94.64%.

	ML_method	accuracy
0	logit	81.25
1	LASSO	81.25
2	support vector machine	94.64
3	Random Forest	85.71
4	GradientBoostingClassifier	87.50
5	k-nearest neighbors algorithm	92.86