INTERNATIONAL JOURNAL OF RECENT TRENDS IN TECHNOLOGY AND ENGINEERING (IJRTTE)

Journal Home Page: https://ijrtte.com/

IJRTTE

Lung Cancer Prediction Using Machine Learning

Migavel M [1] and Privietha P [2]

[1]MCA Student, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India.
[1]migavelmigavel2001@gmail.com, [2]priviethaprabhakar@gmail.com

Abstract. Lung cancer is a type of cancer that starts in the lungs and cannot be prevented at the final stage but its risk can be reduced by treatments. Therefore, detection of lung cancer at the early stage is possible to reduce survival rate. The number of chain smokers is probably equal to the number of people affected by lung cancer. The lung cancer is predicted using the Logistic Regression. The study uses a logistic regression for categorical datasets. The model is obtained after parameter assessment, test the significance of each affecting attribute, and test the model. This is done to obtain prediction models and risk factors at the level of correlation of disease size. The results using logistics regression model for prediction of lung cancer patients based on symptoms, habits, and history of health diseases etc. to see the level of risk could have lung cancer. Some of the symptoms that affect a person with lung cancer are smoking, drinking alcohol, difficulty swallowing, coughing, chronic diseases, fatigue, and age.

Keywords: Prediction, Logistic Regression, Machine Learning.

1. Introduction

Lung cancer is considered one of the deadliest diseases in the world today. Lung cancer affects people more widely presently ranks high in the death. Lung cancer is found in the lung and there are two types of lung cancer.

One is non-small cell and the other is tiny lung cancer of the cell. Some of the common health problems for patients include chest pain, respiratory disease, weight loss, dry cough, etc. smoking and second-hand cigarettes are the main causes of lung cancer. Lung cancer treatment involves operations, immune therapy, chemotherapy, radiation therapy, etc. Despite this technique of diagnosing lung cancer, the clinician may only know this at an advanced Nat. Early precaution has more survival rate than the last stage, in order to prevent the dead rate low. Survey on the cancer survival are very high even after correct treatment and diagnosis. Lung cancer survival rates different from person to person. It depends on age, gender, race and health. In first stage of lung cancer, this prediction model plays an important role in detecting and predicting the lung cancer of the patients.

1.1 Machine Learning

Machine learning is programming computers to optimize a performance criterion using example data or past experience. The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Machine Learning is one of the most discussed topics in research and industries. It focuses on the development of software programs that can process data and use it to learn and make few predictions

Volume: 02 Issue: 01 Year: 2023
Received: 06.01.2023 Received in revised form: 08.02.2023 Accepted: 14.03.2023
Available Online: 31.03.2023
Published by NTL Publisher

NTL Publication

based without being explicitly coded. Machine Learning algorithms are mainly classified into three methods: Supervised Learning, Unsupervised Learning, and Reinforcement Learning

Machine learning is a subfield of AI, which enables a computer system to learn from data. ML algorithms depend on data as they train on information delivered by data science. Without data science, machine-learning algorithms will not work as they train on datasets. No data means no training.

1.2 Logistic regression

Linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination).

Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; [2] the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names

2. Review of Related Literature

The wrapper feature selection method as well as stochastic diffusion research algorithm on lung cancer concluded that this is one of the best performing algorithms for classification. The modified stochastic diffusion (SDS) algorithm, a novel feature selection algorithm based on the wrapper. In order to define optimum subsets of features the SDS benefits from direct contact by agents.[10]

The Northern Centralized Cancer Group (NCCTG), suggested better results in better data between the logistic regression, Naive Bayes, and classifications on lung cancer and predicted that would be better classified because of the increase in lung cancer training data [2]. The novel extraction method for visual data and used machine learning classifications to enhance precision [7]. Suggested to categories 91 percent of correct benign and malignant data as the 3D CNN, supervised learning of the lung nodule and unattended logistic regression approaches [3]. A review of the significance of the neural network of Convolution with a precision of nearly 90% for the prediction of a pulmonary module [5]. Optimization of fluorescent particle swarm by a deep neural network in pictures of lung cancer to reach 99.2% accuracy. [9]

The Most Significant Source of Death for Both Women and Men Is Lung Cancer, A Disease of Uncontrolled Cell Growth in The Lung Tissues [1]. Data Analysis Is A Crucial Role in The Growth of The Discovery of Data in Datasets. It Has Many Potential Uses. The Performance of Classifiers Is Highly Dependent on The Data Set Used for Learning. These Results in Improved Effective Classification Models in Terms of Predictive or Descriptive Accuracy Reduced Computational Time Needed to Construct Models as They Learn More Quickly, And A Greater Understanding of The Models. This Offers A Comparative Study of Data Classification Precision, Using Lung Cancer Data in Various Scenarios. This Compares Predictive Performance of Rising Classifiers in Quantitative Terms. [8]

3. Methodology

The research paper is a machine learning using Logistic regression. The dataset has been taken from online. The data are split into 80% training and 20% testing data. Then logistic regression model has implied and applied to the data and the result of the patient is displayed based on accuracy classification.

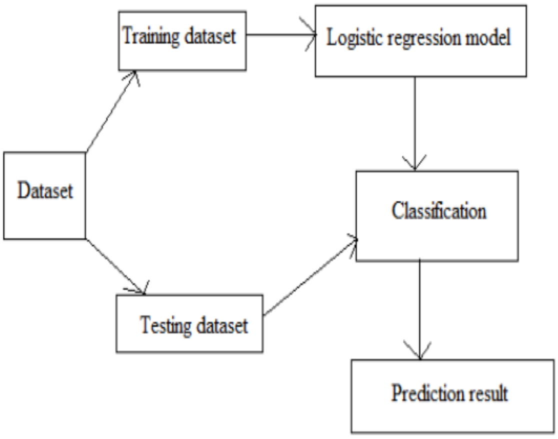


Figure 1: Pictorial Representation of Methodology.

Figure 1 represents the flow of the data in the classification process. Data are split for testing and training and then sent for model classification. Finally, the model is used for prediction.

a) Collection of data:

Most suitable input data has been taken, data can be found from any sources. In this research paper, the datasets are collected from online.

b) Lung cancer dataset as input:

In this step, we give the dataset as an input to the proposed system and it gives the result. The dataset is split into two parts training and testing.

c) Feature selection:

The aim of the feature selection is to identify those inputs, which are related with output values, and the values depend upon some input, which is chosen by using some test.

d) Splitting dataset:

The dataset has 900 test results where the dataset is split into training and testing part. With the training and testing, we can get accuracy score.

e) perform lr technique on training datasets:

Logistic regression is one of the popular models used for analyzing many datasets in the machine learning. In this logistic regression is mainly used for classification purpose.

4. Modeling and Analysis

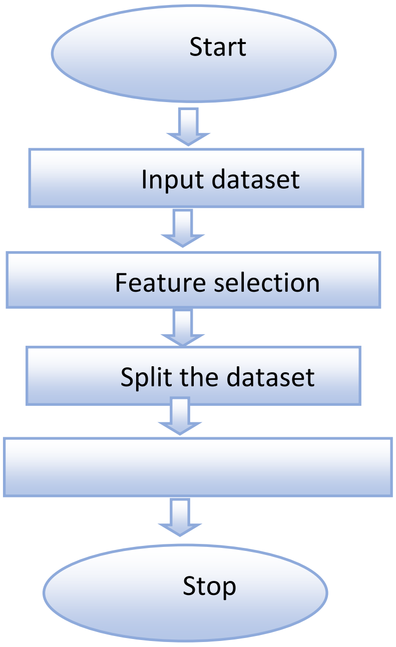


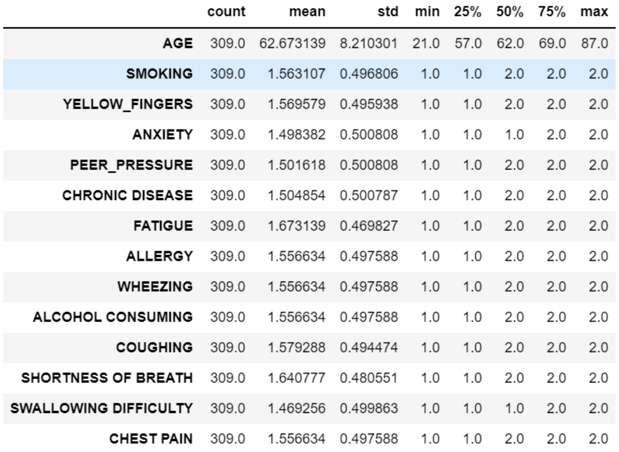
Figure 2: Flow chart for Model Prediction.

The above figure 2 describes the flow of the system where at first it takes the dataset as an input. Then feature selection process is done for getting the expected results. The dataset is split into training and testing dataset and then the logistic regression is used for ANALYZE the datasets. Then performance is classified by using classification model and the above criterion used are accuracy, precision and score.

5. Dataset

The data used in this study was taken from online and this is the data collection from the website of the lung cancer prediction. The dataset consists of 12 attributes and 300 rows of data. The attributes include 'Gender', 'smoking', 'age', 'Anxiety', 'yellow fingers', 'wheezing','Allergy', 'Consuming alcohol', 'Coughing',, 'Chest pain', and 'Lung cancer'. Pre-processing data is done only by checking missing values and there is no missing value in the dataset.

There are many methods to deal with imbalanced classification problems for imbalanced data, some are training data and developing other versions of the existing machine-learning algorithm. To improve the accuracy for an imbalanced dataset is using the logistic model. Data preparation aims to specify independent variables and dependent variables.



	count	mean	std	min	25%	50%	75%	max
AGE	309.0	62.673139	8.210301	21.0	57.0	62.0	69.0	87.0
SMOKING	309.0	1.563107	0.496806	1.0	1.0	2.0	2.0	2.0
YELLOW_FINGERS	309.0	1.569579	0.495938	1.0	1.0	2.0	2.0	2.0
ANXIETY	309.0	1.498382	0.500808	1.0	1.0	1.0	2.0	2.0
PEER_PRESSURE	309.0	1.501618	0.500808	1.0	1.0	2.0	2.0	2.0
CHRONIC DISEASE	309.0	1.504854	0.500787	1.0	1.0	2.0	2.0	2.0
FATIGUE	309.0	1.673139	0.469827	1.0	1.0	2.0	2.0	2.0
ALLERGY	309.0	1.556634	0.497588	1.0	1.0	2.0	2.0	2.0
WHEEZING	309.0	1.556634	0.497588	1.0	1.0	2.0	2.0	2.0
ALCOHOL CONSUMING	309.0	1.556634	0.497588	1.0	1.0	2.0	2.0	2.0
COUGHING	309.0	1.579288	0.494474	1.0	1.0	2.0	2.0	2.0
SHORTNESS OF BREATH	309.0	1.640777	0.480551	1.0	1.0	2.0	2.0	2.0
SWALLOWING DIFFICULTY	309.0	1.469256	0.499863	1.0	1.0	1.0	2.0	2.0
CHEST PAIN	309.0	1.556634	0.497588	1.0	1.0	2.0	2.0	2.0

Figure 3: Representation of Dataset

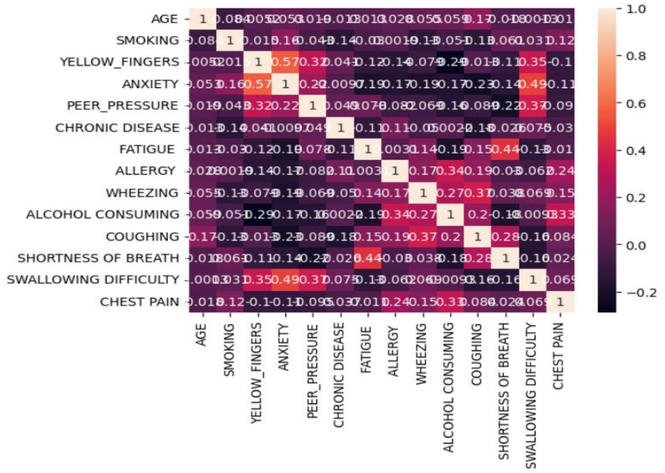


Figure 4: Pictorial Representation of Dataset

Figure 3 and 4 represents the dataset with background variables as symptoms and the mean, standard deviation, minimum value and the maximum value range in the dataset is sorted based on the background variables for future process.

6. Implementation

To implement we need logistic regression equations. The lung cancer dataset included attributes hence the hypothesis for lung cancer detection is of the form.

$$h_\theta(x) = \frac{1}{(1+e^{-\theta^T x})}$$

Where $\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots\ldots + \theta_{15} x_{15}$

Here x1, x2, x3, x15 are 15 attributes which facilitates lung cancer. Total number weights in these case is 16 including s

The cost function for logistic regression is given by

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \ \log(h_\theta(x^i) + \ (1 - y^i \ ) \ (1 - \log(h_\theta(x^i))$$

where m is the total number of input instances.

The objective here is to minimize the cost function parameterized by 0, using gradient descent rule

$\min_\theta \mathrm{J}(\theta)$, where θ_j is computed as follows

$$\theta_j := \theta_j - \alpha \frac{\delta J}{\delta \theta_j} \qquad 0<=j<=15$$

Here partial derivates of cost function parameterized by θ,using gradient descent rule ie..,

$$\frac{\delta J(\theta)}{\delta \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - \ y^i)$$

$$\frac{\delta J(\theta)}{\delta \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - \ y^i)x_j^i \qquad 1<=j<=15$$

Psuedocode:

```
1: Initialize all weights θ 's to zero

2: Use equation (3) to estimate θTx

3: Obtain the hypothesis θ(x) from equation (2)

4: Compute the cost function θ(8) from equation (4)

5: Compute the gradients using equation (6) and equation (7). Update the weights θ 's using (5).

6: Go to step2, repeat the process until the weights do not change.

7: For predicting class of new data, optimal weights corresponding global minimum cost function is recorded and substituted in step3. If probability is above 0.5 lung cancer positive else lung cancer negative.
```

7. Result

In this research work, we predict whether a lung cancer using logistic regression affects the patient. The dataset is divided into two part for training and testing. Using logistic regression, the system is trained with different training dataset such that it can predict lung cancer and then the system is tested using testing dataset for the accurate result. Using that score, we can predict the person is having lung cancer or not.

The testing score of the model is 85.48387096774194.

The Training score of the model is 88.25910931174089.

8. Conclusion

The major and frequent bases of cancer deaths globally in terms of both instance and transience is lung cancer. There are many dead cases because most of the people do not take the proper treatment in the early stage that is why it is hard to cure the patients in the final stage of the lung cancer.

So, taking precaution in the first stage of the lung cancer reduces the causality rate of the lung cancer. The lung cancer can be predicted with the help of the machine learning where the proposed system can predict the lung cancer in the early stages, which helps the causality rate of the patients.

9. Future Enhancement

Instead of prediction from dataset and numerical feature we can use image recognition and deep neural network. We can create a leverage the machine learning model by creating a licensed online portal for predicting lung cancer and lend the software to hospitals.

References

1. Guruprasad Bhat, Vidyadevi G Biradar, H Sarojadevi Nalini, (2012), “Artificial Neural Network based Cancer Cell Classification (ANN – C3)”, Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.
2. Privietha P, Joseph Raj V (2020), “Deep Learning Technic on Gait Analysis” published in Test Engineering and Management, Volume: 83, May –June 2020, SJR:0.1, ISSN: 0193-4120, pp: 11817 -11823.
3. Privietha P, Joseph Raj V (2022), "Hybrid Activation Function in Deep Learning for Gait Analysis," 2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for Sustainable Future (PECCON), Chennai, India, 2022, pp. 1-7, doi: https://doi.org/10.1109/PECCON55017.2022.9851128.
4. Vaishnavi. D, Arya. K. S, Devi Abirami.T,M. N. Kavitha B.E-CSE, Builders Engineering College, Kangayam, Tirupur, Tamil Nadu, India. Department of CSE, Builders Engineering College, Kangayam, Tirupur, Tamil Nadu, India
5. Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, “ Artificial Neural Network based Cancer Cell Classification (ANN – C3)”, Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.
6. Mohamad Sayed, “Biometric Gait Recognition based on machine learning algorithms”. Journal of Computer Science, vol. 14(7), pp.1064 – 1073. DOI: 10.3844/jcssp.2018.1064.1073, 2018.
7. Wu Liu and Cheng Zhang, 2018, ‘Learning Efficient spatial-temporal gait features with deep learning for human identification’, Springer Neuro Informatics, pp. 457–471, doi:10.1007/s12021-018-9362-4
8. https://arxiv.org/pdf/1803.08375.pdf