Lung Cancer Detection Using Machine Learning

Teena Sabu

Department of Computer Applications
Amal Jyothi College of Engineering
Kanjirapalli,Kottayam

teenasabu2022@mca.ajce.in

Mr.Jinson Devis

Asst. Professor Computer Applications
Amal Jyothi College of Engineering
,Kanjirapalli,Kottayam

jinsondevis@amaljyothi.ac.in

***Abstract*-Lung cancer is a medical condition that is affected in the lungs when cancerous cells starts growing inside it. The mortality rate of people has expanded due to the increasing rate of incidence of lung cancer. Lung cancer is a disease where cells in the lungs multiply uncontrollable. Lung cancer cannot be prevented.But due to early prediction the risk can be reduced. So detection of lung cancer at the earliest is difficult for the survival rate of patients. There are many reasons for causing lung cancer,but the most important reasons are alchahol consumption and smoking. The lung cancer prediction was done using classification algorithms such as Decision tree ,SVM and Logistic Regression. The key objective of this paper is the early detection of lung cancer by evaluating the performance of classification algorithms.**

Keywords Machine learning algorithms,decision tree algorithms,support vector machine algorithms,logistic regression algorithms

I INTRODUCTION

Due to lung cancer many people are died, Lung cancer is a type of cancer. Our lungs have two spongy organs in our chest that takes oxygen when we inhale and at the time of exhale carbon dioxide is released. People who smoke have the highest chance of lung cancer when compared to others. The chance of lung cancer increases with the length of time and number of cigarettes smoked. Lung cancer cannot be prevented ,but early detection may help for survival.Mortality rate due to lung cancer is 19.4% .Early detection of lung tumor is done by using many imaging techniques such as Computed Tomography (CT, Chest X-ray and Magnetic Resonance Imaging (MRI). Detection means classifying tumor two classes (i)non-cancerous tumor

(ii)cancerous tumor.

The chance of survival at the advanced stage is less . Manual analysis and diagnosis system can be greatly improved with the implementation of image processing techniques.The early detection helps to take needed necessary actions. Based upon a person's life style the chance of occurring lung cancer can be predicted.The major reasons for lung cancer are alcohol consumption ,smoking and the quality of air where a person leave is also a factor .Based upon these parameters we can predict a person can affect lung cancer or not. Lung cancer origins at the lung and spreads up to brain .lung cancer detection is very weak because doctor will able to know the disease only at the advanced stage.So,in such situations using machine learning we can predict the chance for occurring lung cancer.

II MACHINE LEARNING ALGORITHMS

Machine learning is the technique of analyzing the data. It is a part of artificial inte lligence based on the idea that systems can learn from data, identify patterns and make decisions with very minimum human interaction.Nowadays machine learning helps a lot of fields including medical field,wether forcasting etc.Using machine learning prediction of data becomes easier. The most important aspect of machine learning is the iterative aspect. Historical computations and datas are used to make relaliable, and accurate results. Machine learning is a study of artificial intelligence. The medical field is using the advanced machine learning algorithms to make predictions based on the historic dataset. By using machine learning prediction of diseases and other things are much better and faster. Machine learning is performed based on previous data .There is two types of machine learning algorithms are used for predictions

1.Supervised learning

2. Unsupervised Learning

In supervised learning ,first stage is to train data each and every data is trained first.Every data must be mapped in the supervised learning process. Supervised learning, also known as supervised machine learning, is a subpart of machine learning and artificial intelligence. It is defined by its use of labelled datasets to train algorithms to classify data

DOI: 10.5281/zenodo.6369574
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

or predict outcomes accurately.By using supervised learning we can get much more accurate result.

In unsupervised data there is no previous learning process,there is no mapping occurred in the unsupervised learning . Unsupervised learning, also known as unsupervised machine learning, It uses machine learning algorithms to analyze and classify data. These algorithms discover masked patterns or data collections without the need of humans. Its ability is to discover the same and differences in data .In unsupervised learning the related datas are formed to groups or clusters.Pattern or similarities based classification is occurred in the unsupervised learning.

Supervised Learning is classified in to two types — Regress ion and classification. Regression is when the algorithm learns from the previously labelled datasets and based upon the results it will predicts the output for the new dataset .Examples of Regression algorithm is Linear regression and Logistic regress ion .Clustering and Assosiation are the types of unsupervised learning. Clustering: grouping related datas into a cluster is called clustering.So,datas with more similarities are comes under a group.Only the related datas are comes under a group. In clustering based upon the commanakities datas are categorized**.**Association: It is one of the method used in unsupervised learning, It is used for finding the relationships of variables in a large data set. It determines the set of items that occurs simultaneously in the dataset.

III DECISION TREE ALGORITHM

One of the most frequently used algorithm in machine learning is decision tree algorithm. It is easy to understand and implement. It first chooses best splitter from the given attributes .It is used as the root node of the tree. Algorithm continues until it finds the leaf node. Decision tree creates training model which is used to predict target value or class value in tree representation each internal node belongs to attribute and each leaf node belongs to class label.Decision tree will help to find the best solution from a group of attributes.In a problem there are different solutions ,decision tree will help to find the best outcome.

IV LOGISTIC REGRESSION

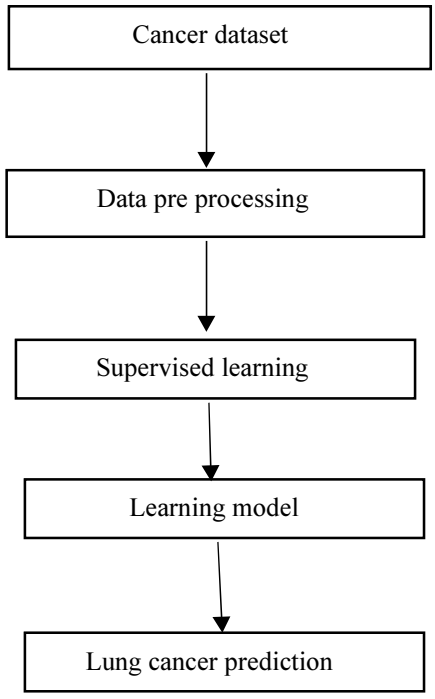
Logistic regression is one of the most broadly used Machine Learning algorithms, it is the part of the supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.It can find the dependent variables from the set of independent variables. It is used to find the result of a categorical dependent variable. So, the output must be a categorical value. Logistic Regression is an important machine learning algorithm because it has the capacity to provide probabilities and categorize new data using continuous and discrete datasets. Logistic Regression can be used to categorize the observations using different types of data and can easily discover the most functional variables used for the classification.

V SUPPORT VECTOR MACHINE ALGORITHM

Support vector machine is another most important algorithm used in machine learning. In support vector machine algorithm each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is known as hyperplane.Hyperplane can divide the data sets.The related values are comes under same place.We can draw hyperplanes at different ways but most perfect hyperplane divides the data set's perfectly. In order to classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vector.The margin should be maximum the we get an accurate result. In linearly supperable data it is possible to divide datas but in non-seperable data set that means the datas are scattered it is not easy to separate the data, to solve this problem support vector machine uses kernel functions which transforms lower dimensional space to higher dimensional space.

VI PROPOSED METHODOLOGY

The aim of this paper is to find wether a person is affected by lung cancer or not.Based upon different parameters it can be evaluated.The structure of the study is as follows



The steps we need to follow are:

1. Collection of data

DOI: 10.5281/zenodo.6369574
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

2. Definig the data
3. Pre- processing data
4. Building model
5. Analysis
6. Results

A data set or dataset is a group of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

The attributes of the dataset used are:

Dataset Description

S.NO	ATTRIBUTES
1	Name
2	Surname
3	Age
4	Smokes
5	AreaQ
6	Alkohol
7	Result

Data Processing

This is the important method in the machine learning method.The lost data and the polluted dataset can decrease the worth of the output that can be formed from the dataset. Data pre-processing can be carried out in order to increase the effect of the data that is retrieved after data processing. In the dataset we can execute data pre-processing by following methods.

1. Removing the missing values- All the arrival that have 0 want to be removed. Because 0 is not a valid entry.

2. Data Splitting- Once the data is cleared it needs to modify the training dataset and test models. once the split data is taken the training data set is used to train the algorithm. A training model is established on the attribute of the training is generated after the training action.

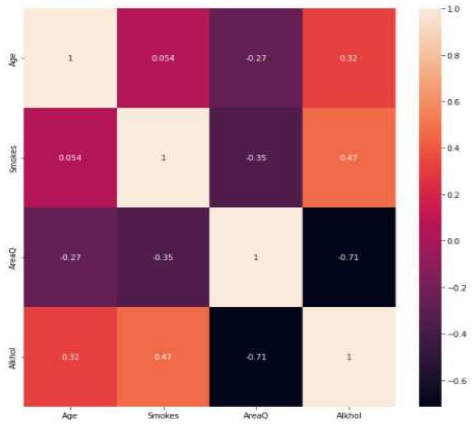


Figure:1

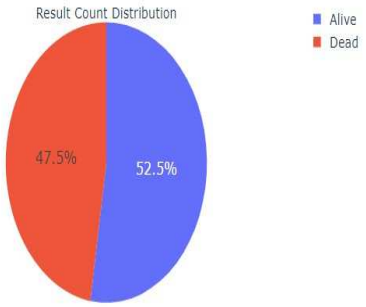


Figure 2

VII RESULTS

We use different algorithms used to classify data and predict the result.By using python programming language implementation is performed.We can create a model for predicting lung cancer based upon different parameters and we can get an accurate result showing how many of the people will cause lung cancer based upon their smoking habbit,alchahol consumption etc.

VIII CONCLUSION

This paper is presented to predict lung cancer based upon the given attributes.Based upon the historical data predictions we get the result.Machine learning algorithms are used to find the results.We can use Decision tree algorithm,Support vector machine algorithm and Logistic regression to find the result.We can different attributes to check the result.Based upon the results we understand that chain smokers have more chance to affect the lung cancer when compared to others.Early prediction of lung cancer may help for the survival.So, my aim is to using this model predict the chance of occurring lung cancer can predict as early as possible IX REFERENCES

[1]https://tlcr.amegroups.com/article/view/21998/16754

[2] https://iopscience.iop.org/article/10.1088/1757899X/1099/1/012059

[3] https://ieeexplore.ieee.org/document/8869001

[4]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6037965/

[5]https://www.tandfonline.com/doi/abs/10.1080/17

DOI: 10.5281/zenodo.6369574
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam