International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
International Conference on
Recent Explorations in Science, Engineering And Technology (ICRESET'19)
Volume-5, Special Issue-March, 2019.

LUNG CANCER PREDICTION USING PREDICTIVE ANALYTICS

Sathish R M.E.,(PhD).[1] Nandhisha S [2] Nivetha M [3] Sudha P [4]

[1,2,3,4]*Department of Information Technology,*

Abstract-Cancer is one of the most dangerous disease that involves unstoppable cell growth to all parts of the body . Lung Cancer is known to be the most highest killer among all type of cancers such as Skin Cancer,Breast Cancer etc. Lung Cancer is also called as Lung Carcinoma, which is a Malignant Tumor described by uncontrolled cell growth in tissues of the Lung . To prevent lung cancer deaths,high risk individuals are screened with low-dose CT scans,because early prediction will double the survival rate of the cancer patients. Usually Lung Cancer is predicted by using MRI (Magnetic Resonance Imaging)scans and CT (Computer Tomography) scans. We proposed a new model to predict Lung Cancer by using textual data. In particular, we investigated about sex,age,Smoking habit,Alcohol consumption, continuous Coughing, Wheezing trouble etc,. In this work,we use Supervised Machine Learning algorithms such as Naive Bayes and Logistic Regression to predict Lung Cancer in terms of accuracy. Aim of our project is to build a model for early prediction of the Lung Cancer which will help the doctor in saving the life of patients.

Keywords-Carcinoma, CT-scan, MRI-scan, K-nearest neighbor

I. INTRODUCTION

Machine Learning is a method of data analysis that automates analytical model building .It is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine Learning enables analysis of massive quantities of data. It may also require additional time and resource to train it properly. Some of the machine learning methods are Supervised learning , Unsupervised learning ,Reinforcement learning.

II. Supervised Machine learning:

It can apply what has been learned in the past to new data using labeled examples to predict future events. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. The Supervised learning which can be categorized based on the Regression and Classification. Algorithms which are widely used such as Naive Bayes ,Support Vector Machine, Random Forests and Decision Tree, Logistic regression ,k-nearest neighbour.

III. Unsupervised machine learning

They are used when the information used to train is either classified or labelled. The system doesn't figure out the right output , but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. There are two types of problems which occurs are as Association and Clustering.The algorithms used in unsupervised learning which are as follows are: K-Means ,Fuzzy clustering,Hierarchical clustering.

IV. Reinforcement learning

It is a learning method that interacts with its environment by producing actions and discovers errors and rewards. Trial and error search are the relevant characteristic of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance.

V . METHODOLOGY:

MACHINE LEARNING PROCESS:

1 Data Preprocessing:

Data preprocessing is one of the most import feature. A huge unstructured data is available, but it is difficult to extract valuable information.While data preprocessing, duplicate values were deleted, there was no missing value that's why missing values technique was not used, steps of data preprocessing are shown in Figure 1.

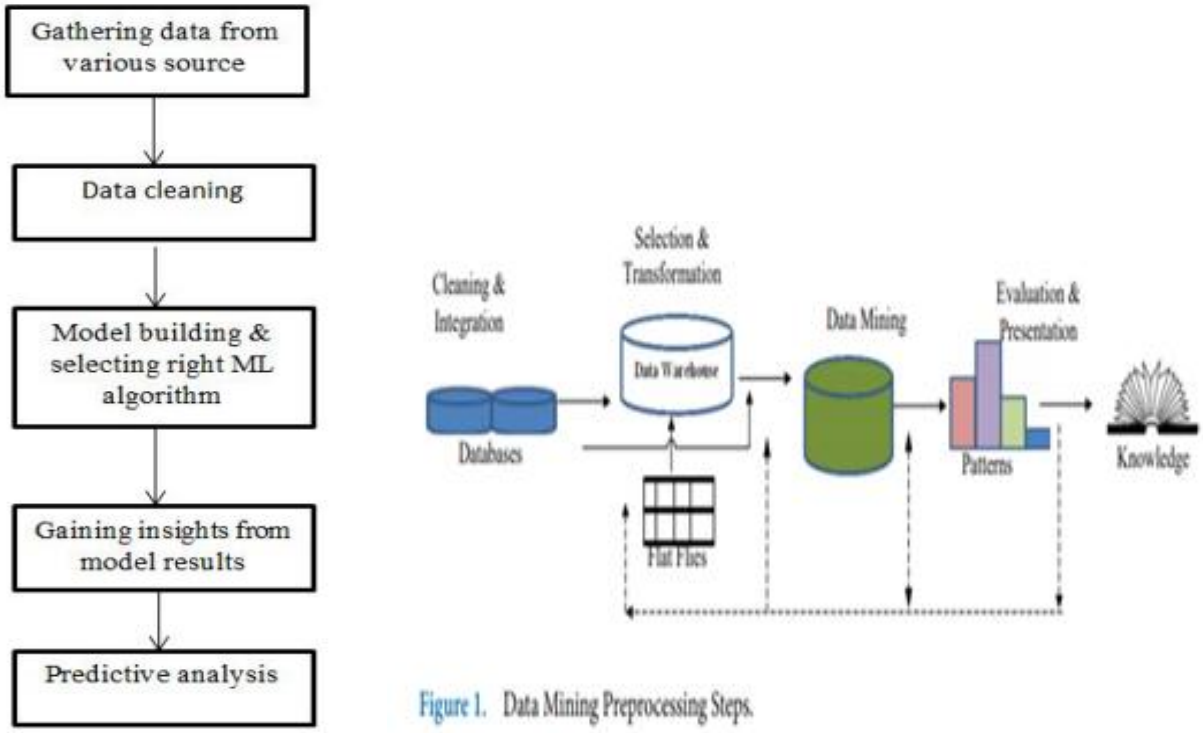


Figure 1. Data Mining Preprocessing Steps.

2 Data Analysis

To get the useful information from the data, data analysis is applied for describe and summaries irrespective of qualitative or quantitative data also identified the relationship/difference between the variables and comparison between the variables.

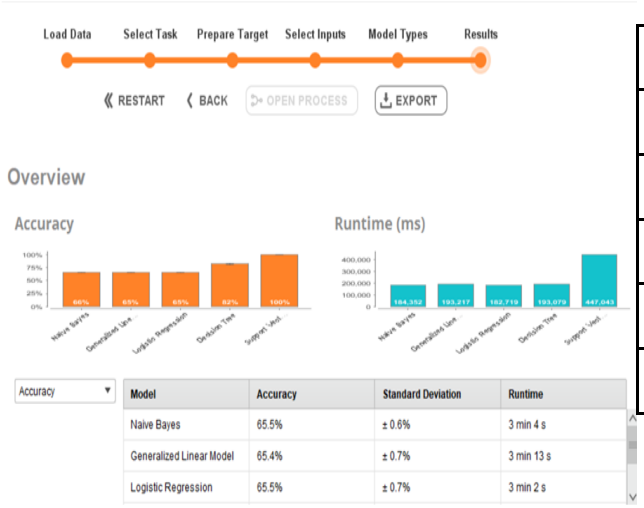
3 Cleaning and Integration

In this stage, irrelevant and unnecessary data was removed from the huge dataset. A clean dataset can give accurate results. Data cleaning phase was conducted before integration in data preprocessing. In data integration data from the multiple sources is combined in a common data source after performing data cleaning steps.

4 Selection and Transformation

In this phase, only relevant data was retrieved by applying feature selection technique and get relevant and decided data. There were large number of attributes in the dataset but keeping in mind the valuable data from the system only interesting attributes were extracted.

VI COMPARISON BETWEEN ALGORITHMS



Algorithm	Accuracy
Naive bayes	65.5%
Decision tree	65.2%
SVM	82.1%
Logistic regression	62%
Linear Regression	51.4%

VII ALGORITHMS

Decision Tree Algorithm

Decision tree is an algorithm used as a support tool for making decisions. [2]It uses a tree-like graph or structure of decisions and their possible outcomes that include the possibilities of an event, resource costs and utility. In a decision tree that has a flowchart-like structure, each internal node is called as a "test" on an attribute (e.g. where a coin flip possible outcomes are head or tail). Each branch refers to the outcome of the test and each leaf node refers to a class label (decision taken after computing all attributes). The path from root to the leaf is called as the classification rules.

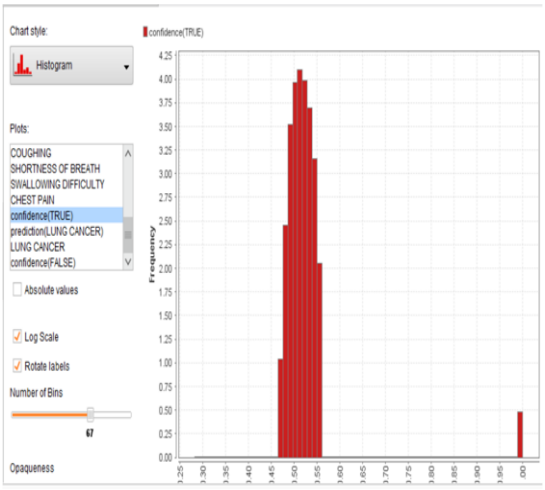
Support Vector Machine (SVM)

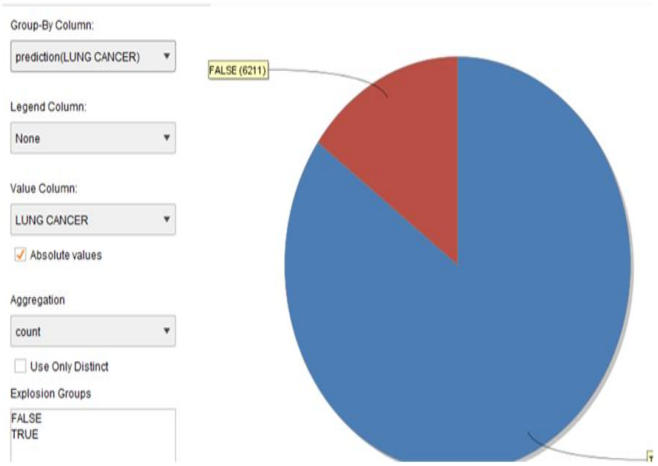
Support Vector Machine (SVM) is mainly used for the classification process.[2] They are built on the idea that it defines the conclusion bordered between groups of instances. A decision plane of SVM is used to separate a set of items from different groups and also distinct a few support vectors in the training set.

Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naïve Bayes is one of algorithm of supervised learning. Naïve Bayes which can be defined as that algorithm that uses Bayes theorem to classify objects. Naive Bayes classifier are a family of simple " probabilistic classifiers" based on Bayes theorem with strong (naïve) independence assumptions between the features. Naive Bayes classifiers are highly scalable ,requiring a number of parameters linear in the number of variables(features/predictors) in a learning problem .It is a simple technique for constructing classifier :models that assign class labels to problem instances, represented as vector of feature values ,where the class labels are drawn from some finite set. Naive Bayes is an conditional probability model; easy to built and particularly useful for very datasets Along with simplicity, Naïve Bayes is known to outperform even highly sophisticated classification methods.Bayes' Rule of conditional probability states that the posterior probability of malignancy given N conditionally independent features F1,... ,Fn and a priori probability,is defined as

$$p(M|F_1,\ldots,F_n)=\prod_{i=1}^{n} p(F_i|M)\frac{p(M)}{p(F_1,\ldots,F_n)}$$





f_measure: 63.34% +/- 0.37% (micro average: 63.35%) (positive class: TRUE)

	true FALSE	true TRUE	class precision
pred. FALSE	1051	1088	49.14%
pred. TRUE	4697	4999	51.56%
class recall	18.28%	82.13%	

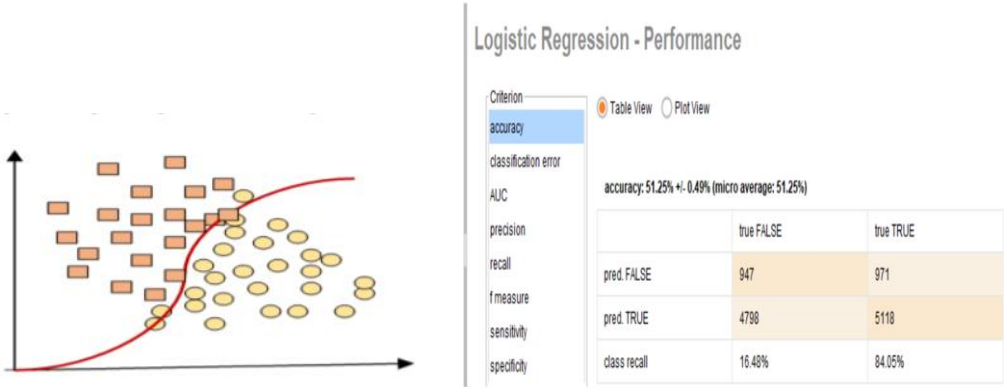
Advantages of Naïve Bayes:

1. It is easy and fast to predict class of test data set .It also perform well in multi class prediction.
2. .When prediction of independence holds ,a Naïve Bayes Classifier performs better compare to other modules like logistic regression and need less training data.
3. It perform well in case of categorical input variables compared to numerical variables.For numerical variable, normal distribution is assumed.
4. It have more runtime when compared to other algorithms so the Naïve Bayes is better.

Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.It is used to describe data and to explain the relationship between one dependent binary variable.The idea of Logistic Regression is to find a relationship between features and probability of particular outcome.

$$p(M|F_1, \dots F_n) = \frac{1}{1+exp[-(\alpha+\sum_{i=1}^{n}\beta_i F_i)]}$$



Logistic Regression classification

VIII DATA SET

Following attributes are used in this paper. [2] The attributes with their description and type, these attributes are used in this research which are used for early detection of lung cancer. These attributes are shown in following.

Table 2:

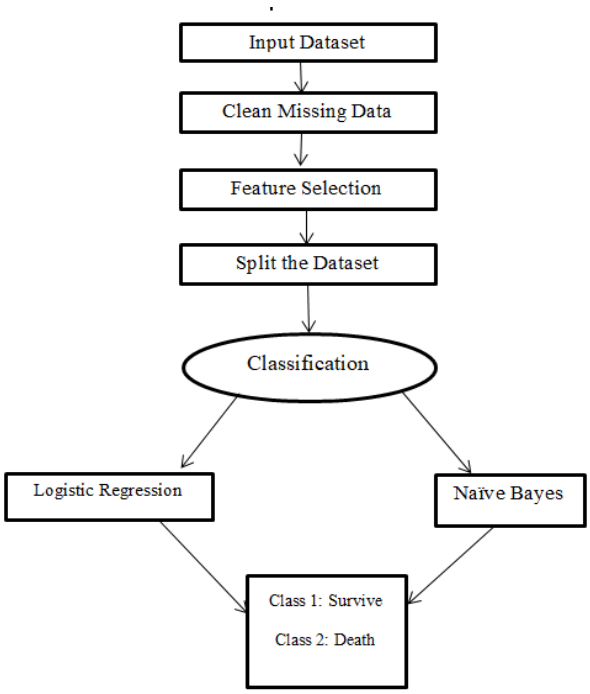
ATTRIBUTE	DESCRIPTION	TYPE
Patient Id	Patients ID	Numerical
Gender	Sex	Numerical
Age	Age in years	Numerical
Smoking	Does patient is smoker	1=yes, 2=no
Yellow fingers	Does patient has yellow finger	1= yes, 2=no
Anxiety	Does patient have anxiety	1= yes, 2=no
Peer Pressure	Does patient have peer pressure	1= yes, 2=no
Chronic disease	Does patient have chronic disease	1= yes, 2=no
Fatigue	Does patient feel tired mentally or physically	1= yes, 2=no
Allergy	Does patient have allergy	1= yes, 2=no
Wheezing	Does patient have wheezing problem	1= yes, 2=no
Alcohol	Does patient consume alcohol	1= yes, 2=no
Coughing	Does patient have cough problem	1= yes, 2=no
Shortness	Does patient have any difficulty in breathing	1= yes, 2=no
Swallow	Does patient have any swallow problem	1= yes, 2=no
Chest Pain	Does patient have chest pain	1= yes, 2=no

IX EXISTING SYSTEM

The lung cancer prediction which are being done through the Chest X-Ray, Magnetic Resonance Imaging (MRI) scan and computed tomography(CT) scans, PET(positron emission Tomography).This scan can sometimes detect disease before it shows up on other imaging tests and Bronchoscopy etc, by the health profession.

X PROPOSED SYSTEM

We proposed a system that will predict lung cancer by using textual data such as age, gender, alcohol consumption, breathing problem ,chest pain etc, We use clustering algorithm to group the data and supervised algorithms to predict the attributes mentioned in data set table.We combine two algorithms such as naive bayes and logistic regression and produce new algorithm that helps us to predict the result with high accuracy.



PSEUDOCODE:

1.Input: get the data from various sources

2. Output: Predictive Model

3.Split dataset into training (70%) and testing data (30%)

4.Identify accuracy of best two supervised algorithm

5. Naive Bayes (dataset)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

6.Logistic Regression

$$f(x) = \frac{L}{1+e^{-k(x-x_0)}}$$

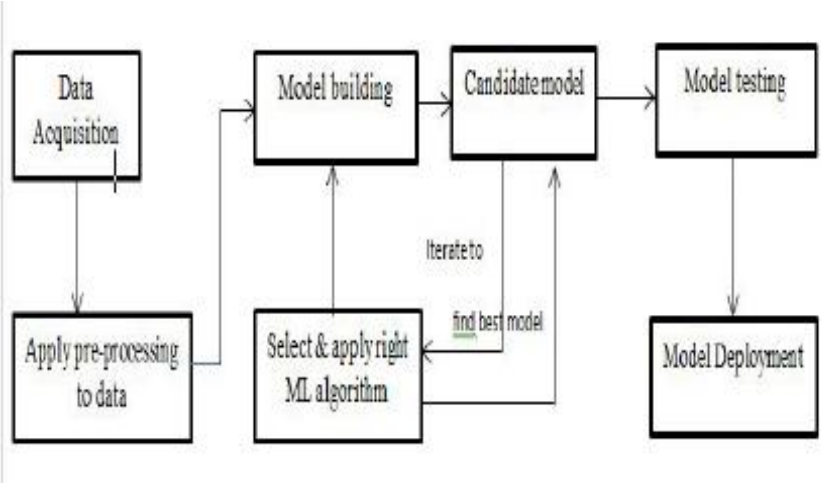
$$y = \frac{e^{b0+b1*x}}{1+e^{b0+b1x}}$$

7.Combine Naïve Bayes and logistic regression to get new algorithm.

8. Apply new algorithm to the model built.

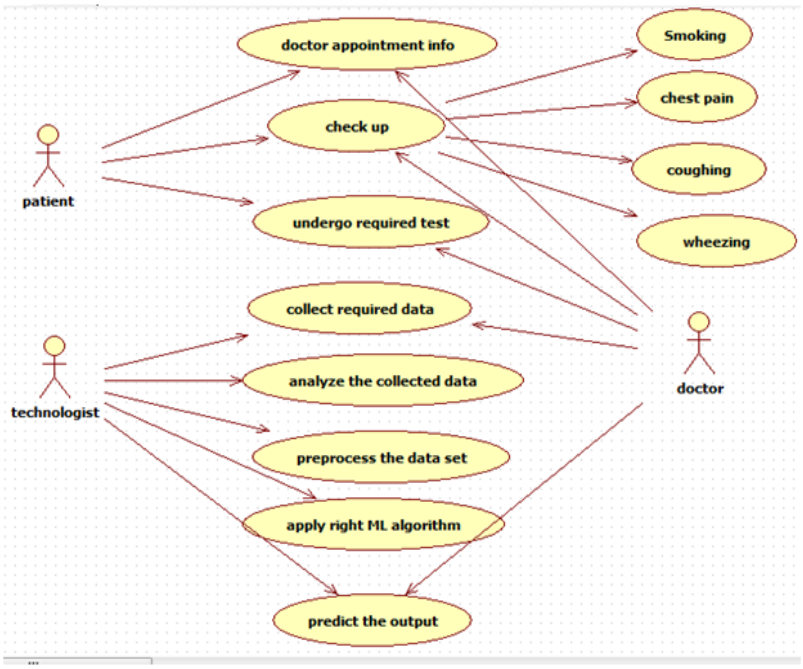
9. Collect real time data to predict accuracy.

ARCHITECTURE DIAGRAM:

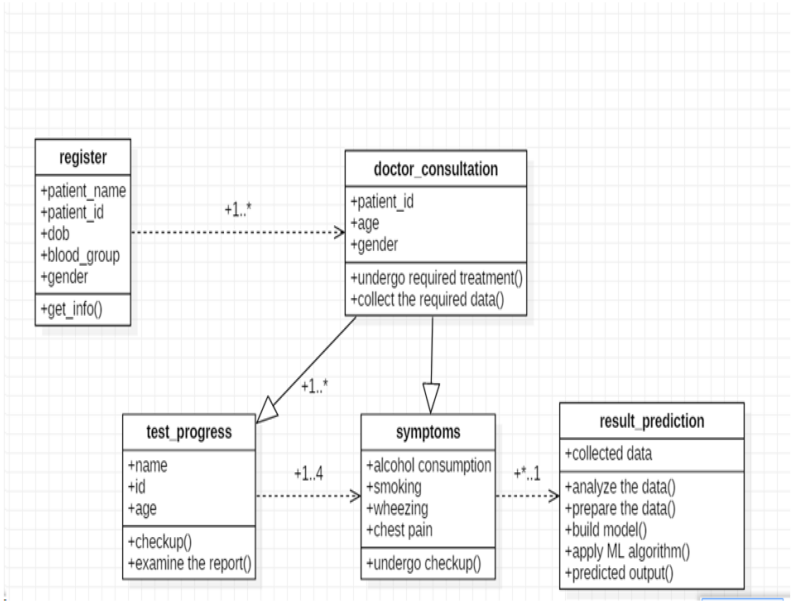


UML DIAGRAMS:

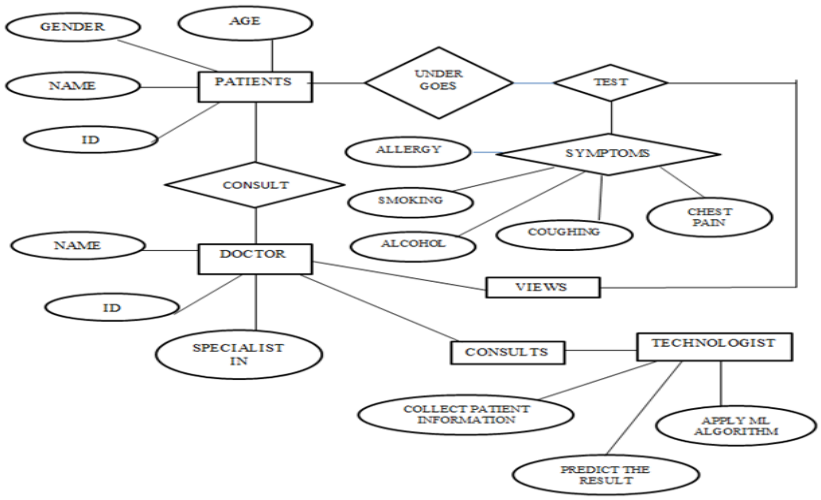
Use Case Diagram:



Class Diagram:



ER diagram:



XI CONCLUSION

Lung cancer is a menacing cancer in the world with high mortality rate. It helps the society to change the lifestyle of the human being to avoid such malignant disease. We have applied data preprocessing techniques on our dataset for removing of noise and dirty data, dirty data is a common term in data mining. Data is cleaned by applying different techniques of data preprocessing in data mining. Classification technique is applied on cleaned data. As shown in Figure 2 that lung cancer ratio is greater in male as compare to female. [2] As shown Figure 3 that age was divided into 5 groups (Group one consists age from 1 to 18, Group Two consists from age 19 to 30, Group Three consists age from 31 to 45, Group Four consists age from 46 to 60 and Group Five consists age 61 to 100. This paper mainly focuses on predicting the lung cancer based on survey using machine learning algorithms. This paper compares various techniques based on efficiency in classification in order to predict lung cancer.

XII REFERENCE:

[1]Survey on lung cancer prediction using ML algorithms by Dr G Vijaya, Er Sathish R , Nandhisha S,Nivetha M,Sudha P in International Journal for Scientific Research & Development in Feb 2019

[2]Survey on lung cancer diagnosis using novel methods by K.Kavitha and Dr.K.Rohini published in International journal of pure and applied mathematics

[3]Classification of Multi-class Microarray Cancer Data Using Ensemble Learning Method published in International journal by B. H. Shekar and Guesh Dagnew

[4]Detection of lung cancer in smokers and non-smokers by applying data mining techniques by Roy Qaiser Hussain and Abdul Azia published in Indian journal of science and technology

[5]Study of classification Algorithm for lung cancer Prediction by Dr .T. Christopher and J.Jamera Banu published in international journal of Innovative science , Engineering and technology

[6]Sex and smoking status effects on the early detection of early lung cancer in high risk smokers by Annette MC Williams,ParmidsaBeigi,Akhila Sriunidhi ,Stephen Lam

[7]Predicting Lung cancer survivability using Ensemble learning Methods by Ali Safiyari and Reza Javidan

[8]Small lung cancer Detection using a Supervised Machine learning Algorithm by Quin Wu and Wenbing Zhao

[9]Real time data collection and analytics of social media sites using Netyltic by R.Sathish and M.Ambika.

[10] Combined Naïve Bayes and logistic regression for quantitative breast sonography Chandra M. Sehgal1 , Theodore W. Cary1 , Alyssa Cwanger , Benjamin J Levenback1 , Santosh S. Venkatesh2 , Departments of Radiology 1, and Department of Electrical Engineering2 , University of Pennsylvania, Philadelphia, PA 19104, USA.