A Review of Lung cancer Prediction System using Data Mining Techniques, Logistic regression, SVM and Naïve Bayes machine learning algorithms

Achyuth SR[1], Apeksha[2], Pradeep K R[3]
[1]*Student, Final year , Dept. of CSE, K.S Institute of Technology, Bangalore, India,*
[2]*Student, Final year Dept. of CSE, K.S Institute of Technology, Bangalore, India,*
[3]*Asst. professor, Dept. of CSE, K.S Institute of Technology, Bangalore, India,*
(E-mail:sr.achu5@gmail.com[1], apekshaaidnal@gmail.com[2], pradeepkr22@gmail.com[3])

***Abstract*—** Cancer has been one of the most important causes of death for both men and women. Detection of the cancer in earlier stages has the high chances of curing the disease completely. Hence the demand for the techniques to detect the occurrence of cancer nodule in early stage is increasing. Treatment and diagnosis of lung cancer in earlier stages can save many lives, failing which may lead to many other severe problems and finally resulting in death of the person. Huge amount of textual data has been collected in healthcare industry but they have not been mined properly to extract the hidden information. Data mining process is a powerful technique that helps to discover patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems. This proposal is used to develop a software which is used to discover the hidden pattern in the lung cancer data set using data mining techniques and machine learning techniques namely Logistic regression (LR), Support Vector Machine (SVM) and Naïve Bayes (NB).

Keywords*—Lung Cancer; Data mining; Logistic Regression; Support Vector Machine; Naïve Bayes.*

INTRODUCTION

Lung cancer is the most frequently causing cancer in both men and women. About 27% of all cancer deaths is from Lung cancer. Lung cancer is the leading cause of cancer death among both men and women[1].Treatment and prognosis depend on the histological type of cancer, the stage, and also performance status of patient. Medications incorporate surgery, chemotherapy, and radiotherapy Survival relies upon organize, general well being, and different elements, yet general just 14% of individuals determined to have lung cancer survive five years after the determination [2]. Mortality and horribleness because of tobacco utilize is extremely high. Normally lung malignancy creates inside the divider or epithelium of the bronchial tree. It can begin at anyplace in the lungs and influence any piece of the respiratory framework. Lung cancer for the most part influences individuals between the ages of 55 and 65 and frequently takes numerous years to create. Lung cancer can be generalized into two subsections, the first one is non-small cell lung cancer(NSCLC) and second one is small cell lung cancer(SCLC). Lung malignancy chiefly happens in elderly individuals. Smoking is the fundamental cause of lung cancer. This includes both the smokers and nonsmokers. Smokers have high possibility of creating lung disease when contrasted with nonsmokers. In the other case lung cancer is identified at a beginning period, the likelihood of cure is high. Essential tumor can be identified early. In the other case that patients don't know essential tumor can develop into metastasis. There are different approaches to recognize the lung malignancy, one of them is to apply its datasets to SVM, NB and LR algorithms and develop the classification and prediction model.

The following symptoms may indicate lung cancer:

- Cough (often with blood)
- Chest pain
- Wheezing
- Weight loss.

These symptoms often don't appear until the cancer is advanced.

People may experience:

- Cough: can be chronic, dry, with phlegm, or with blood.
- Respiratory: frequent respiratory infections, shortness of breath, or wheezing
- Pain areas: in the chest or rib.
- Whole body: fatigue or loss of appetite
- Also common: chest discomfort, hoarseness, or weight loss [3].

Stop the smoking, modification in diet, and chemoprevention can be some of the preventive methods. Screening is a type of auxiliary avoidance. Technique for finding the conceivable Lung cancer patients depends on the efficient investigation of side effects and hazard factors. Non-clinical manifestations and hazard factors are a portion of the non specific markers of the malignancy infections. Natural variables have a critical part in human malignancy. Numerous cancer-causing agents are available noticeable all around we inhale, the nourishment we eat, and the water we drink. The consistent and some of the time unavoidable introduction to natural cancer-causing agents confuses the examination of disease causes in people.

LITERATURE FOR LUNG CANCER

The approach that is being taken after here for the forecast strategy depends on deliberate investigation of the factual components, side effects and hazard factors related with Lung cancer. Non-clinical side effects and hazard factors are a portion of the generic indicators of the growth illnesses. At first the parameters for the pre-conclusion are gathered by collaborating with the obsessive, clinical and medicinal Oncologists.

A. Statistical Incidence Factors

a) Primary histology
b) Age-adjusted rate (ARR)
c) Unrefined occurrence rate
d) Area-related incidence chance[4]

B. Lung cancer symptoms

The following are the generic lung cancer symptoms [5].

- A cough that does not go away and gets worse over time
- Coughing up blood (heamoptysis) or bloody mucus.
- Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
- Clubbing of the fingers and toes. The nails appear to bulge out more than normal.
- Paraneoplastic syndromes which are caused by biologically active substances that are secreted by the tumor.
- Weight loss and loss of appetite
- Increase in volume of sputum
- Wheezing
- Shortness of breath
- Repeated respiratory infections, such as bronchitis or pneumonia
- Repeated problems with pneumonia or bronchitis
- Fatigue and weakness
- New onset of wheezing
- Swelling of the neck and face
- Fever
- Hoarseness of voice
- Puffiness of face
- Loss of appetite
- Nausea and vomiting

C. Lung disease hazard factors

- Smoking Beedi , Cigarette , Hukka.
- Second-hand smoke
- High dosage of ionizing radiation
- Radon presentation
- Word related presentation to mustard gas chloromethyl ether, inorganic arsenic, chromium, nickel, vinyl chloride, radon asbestos
- Air contamination
- Inadequate utilization of natural products and vegetables

DATA MINING

Data mining is the procedure of consequently gathering extensive volumes of information with the target of finding concealed examples and investigating the connections between various kinds of information to create prescient models [6]. The most importance of Data Mining is to find outlines mechanically with smallest customer data and tries. Data Mining is fundamental device equipped for usage decision making and for assessing wants examples of market. In this work, we use the classification techniques. Classification and prediction are two kinds of data examination that can be used to remove models depicting indispensable data classes or to envision future data designs. Such examination can help give us a predominant understanding of the data free to move around at will.[7] Data Mining tools besides, systems can be suitably helpful in different fields in different spaces. Various Organizations now begin using Data Mining as a gainful device, to contract with the commanding surroundings for data examination. By using Data Mining tools and methodology, assorted fields of business get advantage by simply study diverse examples of market and to impact quick and successful market to slant examination. Data mining can be very effective tool to diagnose a disease.

Data Mining Process

Extracting the patterns from the data is the main objective in data mining and making useful application from the data patterns extracted.

Data mining involves following stages [8]

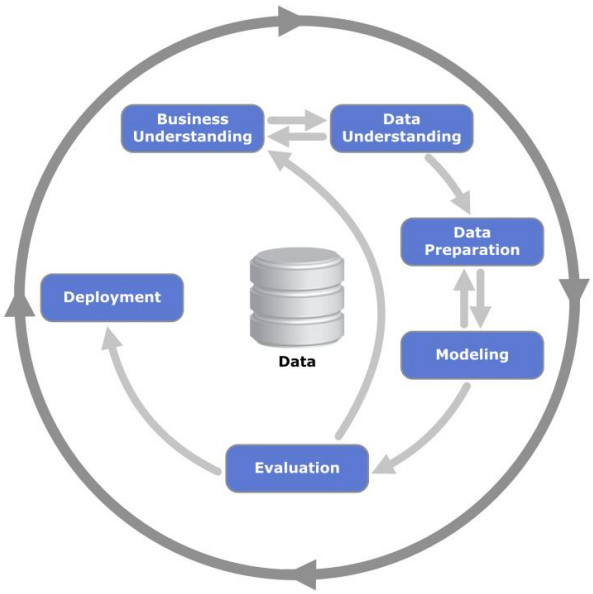

Fig.1 Data Mining Process Representation

D. Problem statement or Business Understanding

Identifying the goals is the initial step in data mining. Depending on the goals defined proper tools should be applied to the data in order to build the corresponding behavioral model.

E. Data exploration

If the data quality is not satisfactory for the designed model then the storage strategies and the future data collection can be reframed for this. For examination, all information must should be solidified so it can be dealt with reliably.

F. Data preparation

The reason for this progression is to clean and change the information, with the goal that absent and invalid esteems are dealt with, and all known legitimate esteems are made predictable for the more vigorous investigation.

G. Modeling

On the basis of data and expected outcomes, a data mining algorithm or combination of algorithms is selected for the purpose of analysis. There are various algorithms from which a particular algorithm can be selected based on the objective to be satisfied and data quality to be particular algorithm can be selected based on the objective to be satisfied and data quality to be analyzed.

H. Evaluation and Deployment

In view of the outcomes the data mining calculations, an investigation is led to decide key conclusions from the examination and make a progression of suggestions for thought.

Techniques

Strategy that are utilized as a bit of Data Mining social event are an unbelievable information mining structure in light of machine learning. Basically assembling is utilized to portray everything in a strategy of information into one of predefined set of classes or of course get-togethers. Demand framework makes numerical frameworks, for example, decision trees, straight programming, neural system and estimations.

METHODOLOGY

WORK FLOW DIAGRAM

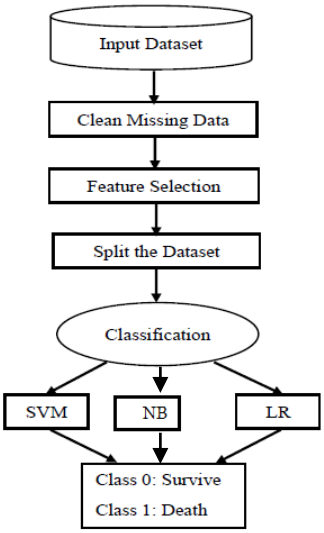

Fig.2 Workflow diagram for predicting lung cancer survivability using support vector machine and logistic regression algorithms respectively.

The Proposed Algorithm for Lung Cancer Prediction

Step 1: Lung cancer data set will be the input for algorithm.
Step 2: Remove the missing values from the dataset.
Step 3: Normalize the dataset.
Step 4: Perform Pearson correlation coefficient (PCC) technique for the purpose of feature selection.
Step 5: Dataset should be divided into 2 parts.
Step 6: Apply the SVM, LR classification and NB techniques on the training dataset.
Step 7: Model should be evaluated.
Step 8: Find and compare the accuracies of the SVM , LR and Naïve Bayes classifiers.[9]

1) SVM: Here, it is utilized for grouping reason. They are based on the possibility of the conclusion level that characterizes conclusion flanked between gatherings of examples. A choice plane of SVM is utilized for partition between an arrangement of things having a diverse gathering of participation and unmistakable a couple of help vectors in the preparation set.

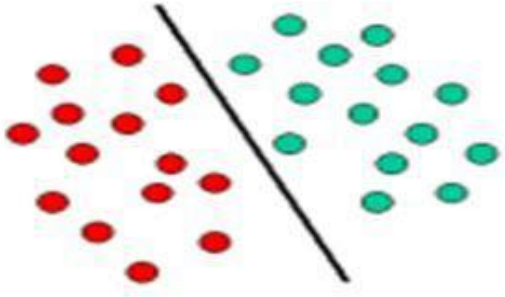

Fig 3: Example of linear support vector machine

Figure 3 gives a predominant case of a direct classifier that separates an arrangement of things into their relating gatherings (green what's more, red for this situation) with a line. Most grouping occupations in spite of the fact that are not that simple and every now and again more composite shapes are required in frame to make an ideal partition i.e. precisely group new experiments on the origin of the examples that train cases [10].

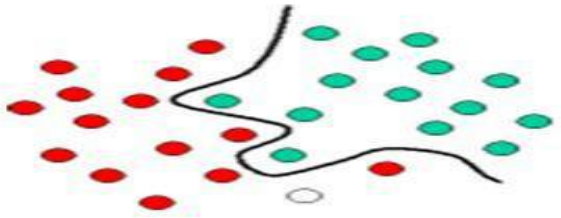

Fig 4: Example of non-linear support vector machine

This circumstance is shown in Figure 4. Characterization employments expand on portrayal isolating lines to separate between things of various gathering enrollments is likewise called hyper plane classifiers. SVM are particularly adjusted to deal with such occupations.

2) Logistic Regression

It is the generalized form of linear regression [11]. Essentially, it is worn for figuring parallel or multi-class subordinate factors. Since the proper reaction variable is discrete, it can't be shown especially by straight fall away from the faith. In two class issue when the result of chances more obvious than half then the class is selected by assigned respect 1 else it is 0. In any case it is a capability perceived pack, it permits that the fitting reaction variable is quick in the coefficients of the figure factors. By at that point, utilizing the experience of information examination the experimenter must pick the primary wellsprings of data and pick their practical relationship to the proper reaction variable.

3) Bayesian Classifier and Naïve Bayes

The Naive Bayes is an energetic procedure for generation of measurable prescient models. NB relies upon the Bayesian hypothesis [12].From a Bayesian perspective, a classification issue can be composed as the issue of finding the class with most extreme likelihood given an arrangement of watched characteristic values. Such likelihood is viewed as the back likelihood of the class given the information, and is typically registered utilizing the Bayes hypothesis. Evaluating this likelihood dispersion from a preparation dataset is a troublesome issue, since it may require a huge dataset to fundamentally investigate all the conceivable mixes. Then again, Naive Bayesian is a straightforward probabilistic classifier in light of Bayesian hypothesis with the (naive) autonomy supposition. In view of that run, utilizing the joint probabilities of test perceptions and classes, the calculation endeavors to evaluate the contingent probabilities of classes given a perception. In spite of its straightforwardness, the Guileless Bayes classifier is a strong strategy, which appears overall great execution regarding arrangement precision, additionally when the freedom suspicion does not hold.

V CONCLUSION

One of the major and incessant bases of cancer deaths all inclusive regarding both example and short life is lung cancer. The primary explanation for the expanding of deaths from it is distinguishing the sickness of late and blames in successful treatment. As needs be, it is smart to decide the survival potential outcomes among the patients. This undertaking uncovers that Logistic Regression classifier gives the highest precision of contrasted with help vector machine classifier and NB. Likewise, the LR classifier gives greatest order exactness concerning each unique classifier. This work can additionally be upgraded by adjusting LR classifier which gives most noteworthy accuracy. With the assistance of machine learning techniques it is extremely hard to analyze the distinctive therapeutic states of a lung disease patient and forecast of conditions are likewise more basic in nature. It is a testing errand in machine learning and information mining fields to build a particular and computationally proficient classifier for therapeutic applications. This can be an awesome future extent of this exploration. For enormous datasets how these characterization calculations carry on, that is another future extent of this undertaking. Additionally the recognizable proof of specific phase of lung disease should be possible in not so distant future. Another prospect of this exploration is the time and space multifaceted nature investigation of various order calculations on medicinal datasets which can be investigated in the expected work.

ACKNOWLEDGMENT

We would like to express our gratitude to R&D , department of CSE, KSIT, Bengaluru for their constant guidance.

REFERENCES

[1] Jiawei Han and MichelineKamber, “Data Mining Concepts and Techniques”, Second Edition, University of Illinois at Urana-Champaign, 2006

[2] "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" V. Krishnaiah et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 - 45.

[3] https://www.gstatic.com/healthricherkp/pdf/lung_cancer_en_IN.pdf

[4] A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM) pp. 2190-2195 © Research India Publications. http://www.ripublication.com

1H Bharathi, 1TS Arulananth ,International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 10 2017

[5] Expert system for detection of breast cancer based on association rules and neural network. Journal: Expert systems with Applications, Murat Karabhatak, M.CevdetInce 2008.

[6] Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com National Conference on Advances in Computer Science and Applications (ACSA-2016) organized by PG and Research Department of Computer Science, Joseph Arts and Science College, Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques 1Vidya R,

2Latha V and 3Venkatesan S 1Assistant Professor, 2,3M.Phil Scholar, 24th Sep 2016.

[7] R.Vidya and G.M nasira "A novel medical support system for the social ecology of cervical cancer:A research to resolve the challenges in pap smear screening and prediction at firm proportion" advances in natural and applied science 9.6 SE 2015.

[8] TheDataminingProcess.[Online].Available:http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html.Shelly Gupta et al./Indian Journal of Computer Science and Engineering (IJCSE).

[9] International Journal of Computer Applications (0975 – 8887) Volume 174 – No.2, September 2017 Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms,AnimeshHazra, NanigopalBera, AvijitMandal

[10] Peng Guan, Desheng Huang, Miao He, and Baosen Zhou, "Lung Cancer Gene Expression Database Analysis Incorporating Prior Knowledge with Support Vector Machine-Based Classification Method", J ExpClin Cancer Res, 28(1): 103, 2009.

[11] Van Belle V, Pelckmans K, Van Huffel S, and Suykens JA, "Support Vector Methods for Survival Analysis: A Comparison Between Ranking and Regression Approaches", ArtifIntell Med, 53(2):107-18, doi: 10.1016/j.artmed.2011.06.006, Aug. 2011.

[12] A survey on predictive analysis of Cancer Survivability Rate using machine learning algorithm Arjun Sharma ,et .al. , Dept. of CSE, K.S Institute of Technology, Bangalore, India.

Achyuth S R pursuing his Bachelor's Degree from KSIT, affiliated to VTU, Belagavi, India

Apeksha pursuing her Bachelor's Degree from KSIT, affiliated to VTU, Belagavi, India

Pradeep K R received his Master's Degree from SJCE, Mysore affiliated to VTU, Belagavi, India and is currently pursuing his PhD from VTU, Belagavi, India. His area of research includes Big Data, Machine learning, Health care analytics