Prediction of Lung Cancer Symptoms Using Naïve Bayes and J48 Classification Techniques

Y. Hemalatha[1] Mr. G. AnanthNath[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of PG Studies, Tirupathi, India

Abstract— Cancer is the disease which is most dangerous that leads to death for both men and women. Lung cancer is uncontrollable disease if they affect both lungs. Early diagnosis of lung cancer saves enormous life's, failing which may lead to another severe problems causing sudden fatal end. The advance detection of cancer is not easier process but if it is detected, it is curable. We analyzed the lung cancer prediction using classification algorithm such as Naive Bayes and J48 algorithm. Initially 100 cancer and non-cancer patients' data were collected pre-processed and analyzed using a classification algorithm for predicting lung cancer. The dataset have 100 instances and 25 attributes. The main aim of this paper is to provide the advance warning to the users and the performance analysis of the classification algorithms. By using classification algorithms Naïve Bayes and j48 .we are going to analyze the lung cancer. Here both cancer and non cancer patient details are collected for processing and analyzing the data which have different instances and attributes.

Key words: Data Mining, Lung Cancer Prediction, Classification, Naive Bayes and J48

I. Introduction

Data Mining is defined as the procedure of extracting information from large sets of data. Data mining is a critical advance in revelation of getting to know from sizeable informational collections. Information mining has observed its essential preserve in every area along with well-being care.Data mining is major role in extracting the hidden information in the medical data base. Mining process is better than the data analysis which includes classification, clustering, association rule mining and prediction. Lung cancer is the most common cause of cancer death worldwide. If the actual lung cancer has spread, a person may feel symptoms in other places in the body. The lung cancer symptom is used to predict risk level of disease. The main aim of this study is predict the risk level of lung cancer using WEKA tool. Symptoms that may suggest lung cancer include:

- Dyspnea (shortness of breath with activity),
- Hemoptysis (coughing up blood),
- wheezing,
- Chest pain or pain in the abdomen,
- Cachexia (weight loss, fatigue, and loss of Appetite),
- Dysphonic (hoarse voice),
- clubbing of the fingernails (uncommon),
- Dysphasia (difficulty swallowing),
- Pain in shoulder, chest, and arm
- Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

Mortality and morbidity receivable to tobacco use is very high. Usually lung cancer develops within the wall or deciduas of the bronchial tree. But it can start each and every in the lungs and affect any part of the respiratory system. Lung cancer mainly affects people between the ages of 55 and 65 and often takes many years to develop.

There are two major types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer grows and spreads in different ways, and is use differently. If the cancer has features of both categories, it is called mixed small cell/large cell cancer. Non-small cell lung cancer is more common than SCLC and it normally grows and spreads more slowly. SCLC is nearly related with smoking and grows more quickly and form large tumours that can spread widely through the body. They often begin in the bronchi near the center of the chest. Lung cancer death rate is similar to total amount of cigarette smoked. Smoking stopping, diet modification, and chemoprevention are primary prevention activities. Screening is a type of elective anticipation. Our method for finding the viable Lung malignant growth sufferers depends on the orderly investigation of symptoms and danger elements. Non-medical manifestations and danger elements are a part of the conventional recommendations of the malignant increase sicknesses. Ecological elements have an important activity in human malignant increase. Numerous cancer-causing agents are to be had sizeable all round we inhale, the sustenance we eat, and the water we drink. The regular and once in a while unavoidable presentation to natural most cancers-inflicting agents muddles the exam of malignancy causes in people. The intricacy of human ailment causes is especially trying for malignant growths with lengthy state of no activity, which are Cancer examine is normally clinical as well as organic in nature, facts driven measurable studies has turned into a normal complement. Foreseeing the result of an infection is a standout among the most interesting and testing errands in which to create statistics mining programs. As the usage of PCs managed with mechanized apparatuses, widespread volumes of medicinal records are being accrued and made accessible to the healing research gatherings. Therefore, Knowledge Discovery in Databases (KDD), which incorporates statistics mining systems, has changed into a widely recognized studies device for medicinal analysts to differentiate and misuse examples and connections among expansive number of things, and made them ready to count on the end result of an illness using the verifiable cases positioned away inner datasets

II. Literature Survey

Krishnaiah V,Narsimha G, Subhash ChandraN [1] proposed to a model for nearly detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Using lung cancer symptoms such as age, sex, wheezing, shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease.

Tapas RanjanBaitharu, Subhendu Kumar Pani [2] Conducted the most important cause of death for both men and women is the cancer lung cancer is a disease of uncontrolled cell growth in tissues of the lung. Data classification is a main task in KDD (knowledge discovery in databases) process. It has several potential applications. The performance of classifiers is strongly reliant on the data set used for learning. It leads to better performance of the classification models in terms of their predictive accuracy, diminishing of computing time needed to build models as they learn faster, and better understanding of the models. A comparative analysis of data classification accuracy using lung cancer data in different plan is presented. The predictive performances of suitable classifiers are compared quantitatively.

The approach that is being followed here for the prediction technique is based on systematic study of the statistical factors, accuracy, execution time, symptoms and risk factors associated with Lung cancer. Non-clinical symptoms and risk factors are some of the common indicators of the cancer diseases. Initially the parameters for the pre-diagnosis are collected by connecting with the pathological, clinical and medical oncologists (Domain experts).

A. *Lung cancer symptoms:*

1) Coughing up blood (heamoptysis) or bloody mucus
2) Weight loss and loss of appetite
3) Wheezing
4) Shortness of breath

B. *Lung cancer risk factors:*

1) *Smoking*
    1) Beedi
    2) Cigarette
    3) Hukka
2) *Second-hand smoke*
3) *High dose of ionizing radiation*
4) *Radon exposure*
5) *Air pollution*

III. DATA MINING CLASSIFICATION METHODS

The data mining consists of various methods. Different methods give different purposes, each method offering its own advantages and disadvantages. In data mining, classification is one of the most essential tasks. It maps the data in to predefined goal. It is a supervised learning as targets are predefined.

The aim of the classification is to build a classifier based on some cases with some attributes to detail the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The most used classification algorithms exploited in the microarray analysis belong to two categories: Naïve Bayes and J48 algorithms.

A. *Data Mining Technique*

Data mining is the process of automatically collecting large volumes of data with the objective of finding hidden patterns and analyzing the relationships between numerous types of data to develop predictive models. The classification techniques and prediction are two forms of data analysis that can be used to extract models describing essential data classes or to predict future data trends. Such analysis can help arrange us with a better understanding of the data at large.

B. *Dataset Description*

Dataset utilized in this investigation is progressively exact and precise so as to enhance the prescient exactness of information mining calculations. Traits for manifestation is utilized to analysis of illness are to be dealt with productively to get the ideal result from the information mining process. The characteristic, for example, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, detached smoker, chest torment, hacking of blood, Fatigue, weight reduction, shortness of breath, wheezing, gulping trouble, clubbing of finger nails, Frequent Cold, Dry Cough, Snoring are taken to consider for anticipating the lung malignancy. WEKA executes calculations for information pre-handling, highlight decrease, characterization, for example, Naive Bayes, J48. The exhibitions of the calculations for lung malignancy ailment are examined utilizing representation instruments.

IV. PROPOSED SYSTEM

Every hospital maintains a huge volume of patients' data. It is very difficult process for analyzing all these records manually. Data mining techniques are used to extract the useful information from the dataset contain huge amount of data. In the healthcare field it helps to analysis the patient's information for provide a warning to the patient who have maximum possibility of chances to be affected with the disease and support the doctors to give a possible treatment for diagnosing. In this work, Naïve Bayes and j48 classification data mining technique is used to find out the most risk factors of congenital Lung cancer. Naive Bayes and J48 algorithms. The naïve Bayes algorithm is an intuitive method that uses the conditional probabilities of each attribute belonging to each class to make a prediction.J48 is an open source java implementation of simple C4.5 decision tree algorithm.J48 is an extension of ID3.The additional features of j48 are accounting of missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. The figure 1 shows the step by step process of the proposed work. This research work predicts the risk factors of congenital lung cancer affecting majority of the patients with the aid of Weka data mining tool.

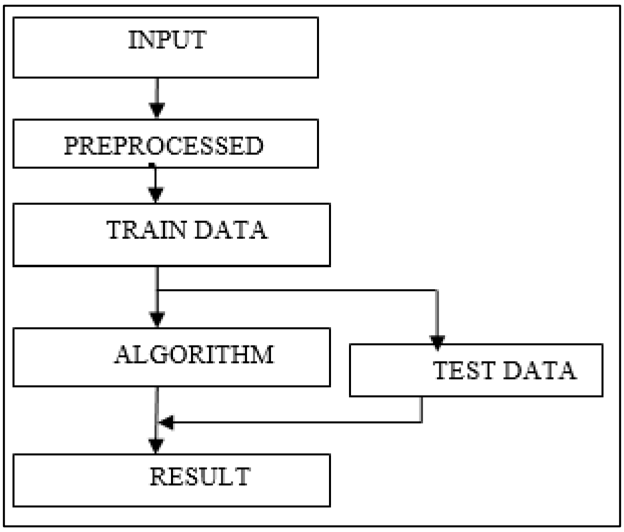


Fig. 1: Proposed Model

A. *Performance Analysis*

The classification algorithms such as Naive Bayes and J48 algorithm is used for predicting the Lung Cancer Symptoms from the given data set instances and the proposed algorithms are applied on type Lung Cancer Symptoms dataset in the WEKA tool and the performance is measured.

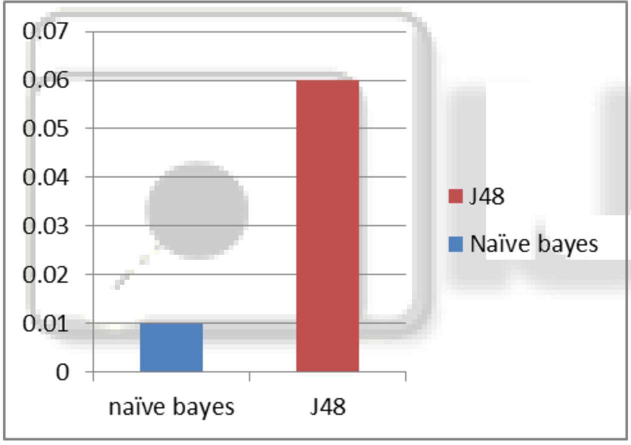


Fig. 2: Comparative analysis

V. PROPOSED ALGORITHMS

A. *Naive Bayes*

Naive Bayes is a classification algorithm for twofold (two-class) and multi-class classification problems. The system is least disturbing to recognize while portrayed using paired or clear cut data esteems. It is called Bayes in mild of the reality that the figuring of the chances for each idea is rearranged to make their computation tractable. As against endeavoring to compute the estimations of each feature), they may be idea to be restrictively free given the and calculated as

P (d1|h) * P (d2|H) and so on.

This is an incredibly strong supposition that is maximum not possible in authentic records, i.e., the characteristics do not interact. All things considered, the methodology performs shockingly nicely on statistics in which this supposition does not preserve.

```
Input:
        Training dataset T,
        F= ($f_1$, $f_2$, $f_3$,.., $f_n$)    // value of the predictor variable
        in testing dataset.
Output:
        A class of testing dataset.
Step:
    1.  Read the training dataset T;
    2.  Calculate the mean and standard deviation of the
        predictor variables in each class;
    3.  Repeat
            Calculate the probability of $f_i$ using the gauss
            density equation in each class;
        Until the probability of all predictor variables ($f_1$, $f_2$,
        $f_3$,.., $f_n$) has been calculated.
    4.  Calculate the likelihood for each class;
    5.  Get the greatest likelihood;
```

The Naïve Bayes model identifies the physical characteristics and features of patients suffering from lung cancer .For each input it gives the possibility of attribute for the expectable state. The Figure 3 shows the implementation of naïve Bayes algorithm on symptoms data.

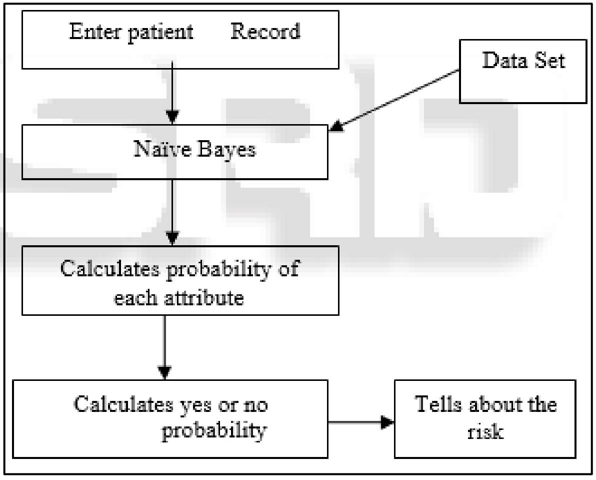

Enter patient Record
Data Set
Naïve Bayes
Calculates probability of each attribute
Calculates yes or no probability
Tells about the risk

Fig. 3: Implementation of Naïve Bayes algorithm on patient data

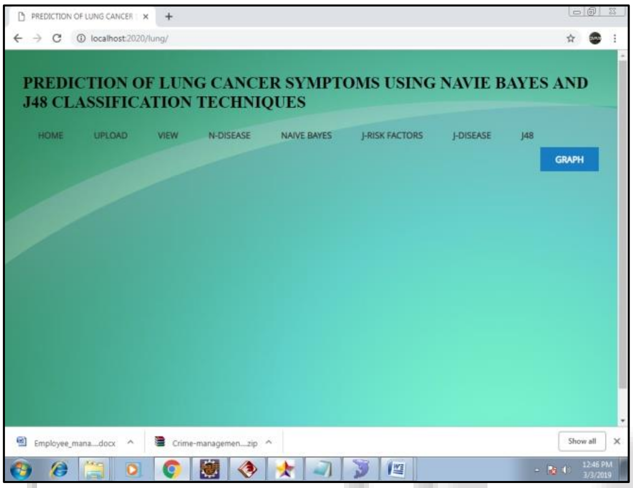
B. *J48 Decision Tree*

Classification is the process of building a model of classes from a set of records and class labels. Decision Tree Algorithm is to find out the manner the features vector acts for various cases. Additionally at the bases of the instruction cases the lessons for the lately produced instances are being observed. This calculation produces the tenets for the expectancy of the objective variable. With the help of tree order calculation the basic conveyance of the facts is results easily reasonable. J48 is a diffusion of ID3. The more highlights of J48 are representing lacking traits, preference timber pruning, ceaseless trait esteem ranges, induction of ideas, and so on.

C. Basic Steps in the Algorithm

1) In the event that the activities have a place with a similar magnificence the tree speaks to a leaf so the leaf is returned by using marking with a comparable elegance.
2) The capacity facts is determined for every trait, given by means of a check on the nice. At that factor the advantage in statistics is decided that would result from a test on the property.
3) Then the first-class first-rate is found primarily based on the prevailing dedication measure and that trait selected for stretching.

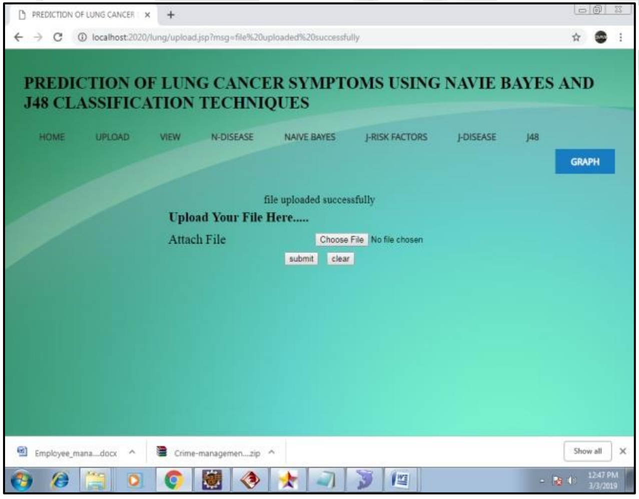
VI. RESULT & ANALYSIS



PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

Fig. 4: Home page

A. Upload



PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

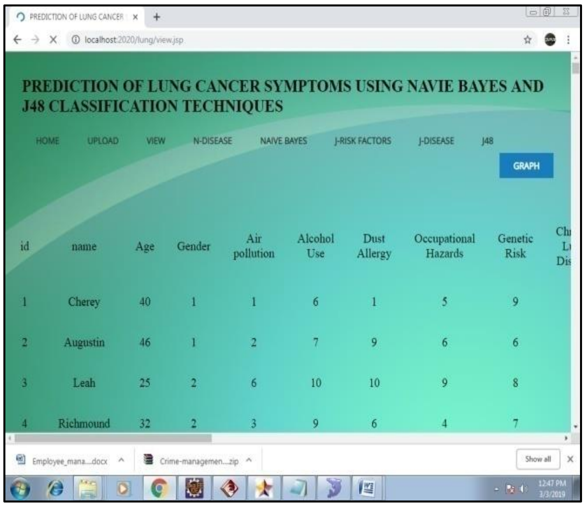
file uploaded successfully

Upload Your File Here.....

Attach File

Fig. 5: Browse the data set for the classification purpose record to be insert the database

B. VIEW

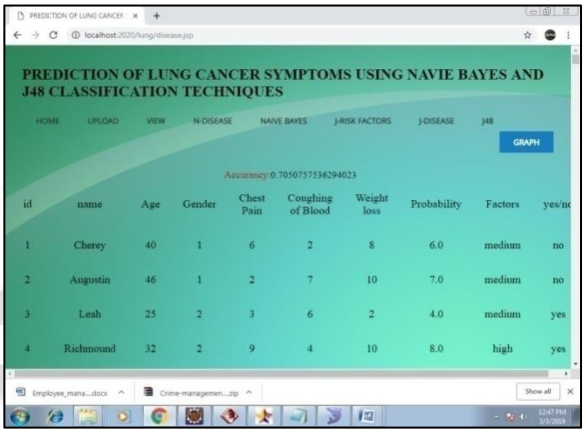


PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

id	name	Age	Gender	Air pollution	Alcohol Use	Dust Allergy	Occupational Hazards	Genetic Risk	Ch L Dis
1	Cherey	40	1	1	6	1	5	9	
2	Augustin	46	1	2	7	9	6	6	
3	Leah	25	2	6	10	10	9	8	
4	Richmound	32	2	3	9	6	4	7	

Fig. 6: To view the dataset and pre-processing the dataset for cancer prediction

C. Naive Bayes-Disease



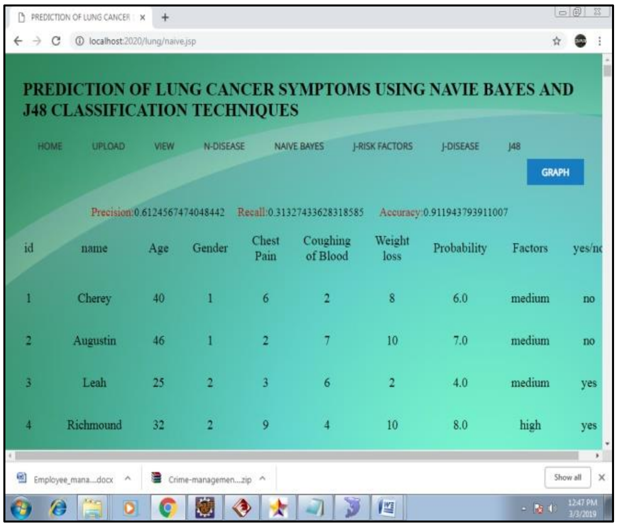
PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

Accurancy:0.7050757536294023

id	name	Age	Gender	Chest Pain	Coughing of Blood	Weight loss	Probability	Factors	yes/no
1	Cherey	40	1	6	2	8	6.0	medium	no
2	Augustin	46	1	2	7	10	7.0	medium	no
3	Leah	25	2	3	6	2	4.0	medium	yes
4	Richmound	32	2	9	4	10	8.0	high	yes

Fig. 7: Prediction accuracy for Naïve Bayes Disease

D. Naïve Bayes



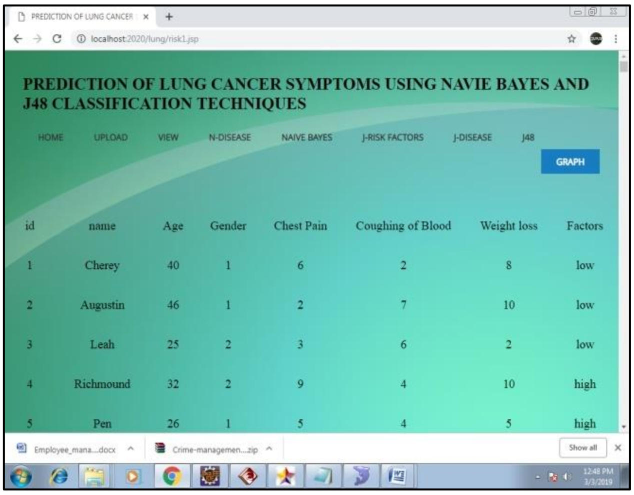
PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

Precision:0.6124567474048442 Recall:0.31327433628318585 Accuracy:0.911943793911007

id	name	Age	Gender	Chest Pain	Coughing of Blood	Weight loss	Probability	Factors	yes/no
1	Cherey	40	1	6	2	8	6.0	medium	no
2	Augustin	46	1	2	7	10	7.0	medium	no
3	Leah	25	2	3	6	2	4.0	medium	yes
4	Richmound	32	2	9	4	10	8.0	high	yes

Fig. 8: Using Naïve Bayes algorithm to find prediction and accuracy on patient details

E. J48-Risk Factors

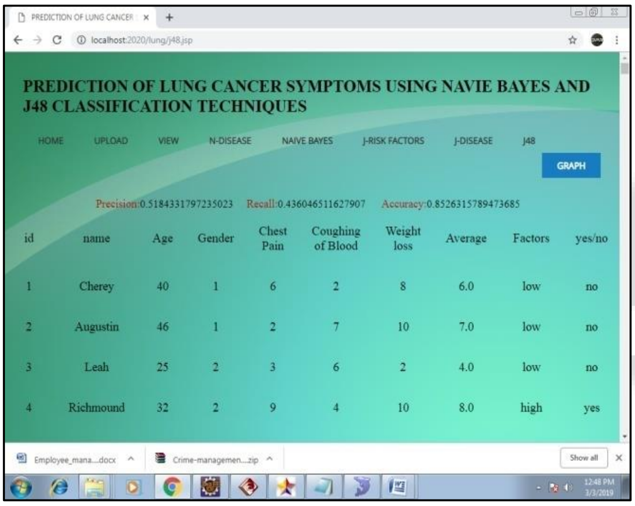


PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

id	name	Age	Gender	Chest Pain	Coughing of Blood	Weight loss	Factors
1	Cherey	40	1	6	2	8	low
2	Augustin	46	1	2	7	10	low
3	Leah	25	2	3	6	2	low
4	Richmound	32	2	9	4	10	high
5	Pen	26	1	5	4	5	high

Fig. 9: Using J48 algorithm to find the risk factors for prediction of lung cancer patent details

F. J48



PREDICTION OF LUNG CANCER SYMPTOMS USING NAVIE BAYES AND J48 CLASSIFICATION TECHNIQUES

Precision:0.5184331797235023 Recall:0.436046511627907 Accuracy:0.8526315789473685

id	name	Age	Gender	Chest Pain	Coughing of Blood	Weight loss	Average	Factors	yes/no
1	Cherey	40	1	6	2	8	6.0	low	no
2	Augustin	46	1	2	7	10	7.0	low	no
3	Leah	25	2	3	6	2	4.0	low	no
4	Richmound	32	2	9	4	10	8.0	high	yes

Fig. 10: Using J48 algorithms to find prediction and accuracy on patient details

VII. Conclusion & Future Scope

Prevalence of Lung cancer disease is high in India, especially in rural areas, which did not get noticed at the early stage; because of the lack of recognition of symptoms. Additionally it isn't always viable for the intentional agencies to do the screening for each considered one of the general populations. The accentuation of this work is to discover the goal amassing of those who desire in addition screening for Lung disease illness, so the frequency and death price may be cut down. Lung infectious increase forecast framework may be additionally upgraded and extended. It can likewise consolidate other information mining methods, e.g., Time Series, Clustering and Association Rules. Consistent statistics can likewise be utilized as opposed to surely absolute information.

References

[1] Krishnaiah, V., Narsimha,G.,SubhashChandra,N., Diagnosis of Lung 48 IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016. ISSN 2348 – 7968 Cancer Prediction System Using Data Mining Classification Techniques, etal,/(IJCSIT) International Journal ofComputer Science and InformationTechnologies, Vol. 4 (1) , 2013. .

[2] Tapas RanjanBaitharu, Subhendu Kumar Pani A, Comparative Study of Data Mining Classification Techniques using Lung Cancer Data, International Journal of Computer Trends and Technology (IJCTT) – volume 22 Number 2–April 2015.

[3] MaryKirubaRani.V,SafishMary.M, Predicting Progression of Primary Stage Cancer to Secondary Stage Using Decision Tree Algorithm International Journal of Advanced Information Science and Technology (IJAIST) Vol.26, No26, June 2014.

[4] Sowmiya.T, Gopi.M, Thomas Robinson, Optimization of Lung Cancer Using Modern Data Mining Techniques ,International Journal of Engineering Research Volume No.3, Issue No.5.

[5] T.Priyanga.A, Prakasam.S, Effectiveness of Data Mining- based Cancer Prediction System (DMBCPS), International Journal of Computer Applications Volume 83 – No 10, December 2013