IPASJ International Journal of Computer Science (IIJCS)
A Publisher for Research Motivation ....
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
Email:editoriijcs@ipasj.org
Volume 5, Issue 11, November 2017
ISSN 2321-5992

CLINICAL DECISION SUPPORT SYSTEM FOR MINING LUNG CANCER DATA USING DATA MINING TECHNIQUES

[1]Padmini.P, [2]Mythili.K
[1]M.Phil Scholar, [2]Assistant Professor
Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram

ABSTRACT

Lung cancer, the foremost cause of cancer associated humanity for together men and women and its occurrence is mounting worldwide. Lung cancer is the unrestrained increase of irregular cells that begin off in one or both Lung. The previous revealing of cancer is uneasy procedure but if it is noticed, it is curable. The major endeavor of this paper is to offer the earliest forewarning to the patient/doctors and the functioning investigation of combining the classification algorithms Naive Bayes and j48. Those Data mining classification algorithms can assist in the prediction of lung cancer research and it improves the quality of healthcare of patients who are affected by lung cancer.

Keywords: Data mining, classification techniques, naïve Bayes, J48.

1. INTRODUCTION

In the existing system is to find out the medical issues of Lung cancer and find out the stages of the lung cancer patients by using the data of Patients Details and risk factor of lung cancer which are collected from the hospital database. The stage of lung cancer refers to the extent to which the cancer has spread in the body. Overall, 10- 15% of lung cancers occur in non-smokers. (Another 50% occur in former smokers).Two-thirds of the non-smokers who get lung cancer are women, and 20% of lung cancers in women occur in individuals who have never smoked. Originally cancer and non-cancer patients' data were composed preprocessed and investigated using a classification algorithm for predicting lung cancer.

2. RELATED WORKS

1. "Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, 2013.

This paper helps to study of how to prevent the lung cancer. First we gather the data from hospitals, data centers and cancer research centers. The collected data is pre-processed and stored in the knowledge base to build the model. To Give Risk scores for the attributes that represent the significant patterns using Decision - Tree algorithm and the data is clustered using K-means clustering algorithm to separate cancer and non cancer patient data based on the risk score. If the patient contains cancer, Test the data and find the risk status using prediction.

2. V. Krishnaiah, Dr. G. Narsimha, R. N. Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies(IJCSIT), 2013 - This paper helps to study of early detection of cancer can be helpful in curing the disease completely., such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data.

3. T.Karthikeyan, P.Thangaraju, "Analysis of Classification Algorithms Applied to Hepatitis Patients", International Journal of Computer Applications, 2013 - This paper mainly deals with various classification algorithms namely, Bayes.NaiveBayes, Bayes.BayesNet, Bayes. NaiveBayesUpdatable, J48, Randomforest, and Multi

Layer Perception. It analyzes the hepatitis patients from the UC Irvine machine learning repository. The results of the classification model are accuracy and time.

4. T.Karthikeyan, P.Thangaraju, "PCA-NB Algorithm to Enhance the Predictive Accuracy" International Journal of Engineering and Technology (IJET), 2014 - This paper mainly deals with feature extraction algorithm used to improve the predicted accuracy of the classification. This paper applies with Principal Component analysis as a feature evaluator and ranker for searching method. Naive Bayes algorithm is used as a classification algorithm. It analyzes the hepatitis patients from the UC Irvine machine learning repository. The results of the classification model are accuracy and time.

5. Chinnappan Ravinder Singh, Kandasamy Kathiresan, "Molecular understanding of lung cancers-A review" Centre of Advanced Study in Marine Biology, Faculty of Marine Sciences 2014 - The purpose of this paper is to review scientific evidence, particularly epidemiologic evidence of overall lung cancer burden in the world. Molecular understanding of lung cancer at various levels by dominant and suppressor ontogenesis.

3. EXISTING MODEL

In the existing model is to classify only by using the x-ray, CT scan for detect lung cancer. The stage of lung cancer refers to the extent to which the cancer has spread in the body. Overall, 10- 15% of lung cancers occur in non-smokers. (Another 50% occur in former smokers). Two-thirds of the non-smokers who get lung cancer are women, and 20% of lung cancers in women occur in individuals who have never smoked. The existing system is limited to find the medical issues of Lung cancer, the stages of lung cancer patients by using the Patients history and risk factor of lung cancer which are collected from the hospital database.

4. PROPOSED MODEL

Lung cancer is the unrestrained increase of irregular cells that begin off in one or both Lung. The previous revealing of cancer is uneasy procedure but if it is noticed, it is curable. The prediction of lung cancer research and it improves the quality of healthcare of patients who are affected by lung cancer. Lung cancer is the unrestrained increase of irregular cells that begin off in one or both Lung. The drug is not effective in curing malaria. The investigation is performed based on combining the data mining classification algorithms such as Naive Bayes and j48 using weka tool. The results thus obtained illustrated with the proposed component designed using ASP.NET which is capable of predicting the lung cancer effectively. Thus, The major endeavor of this paper is to offer the earliest forewarning to the patient/doctors

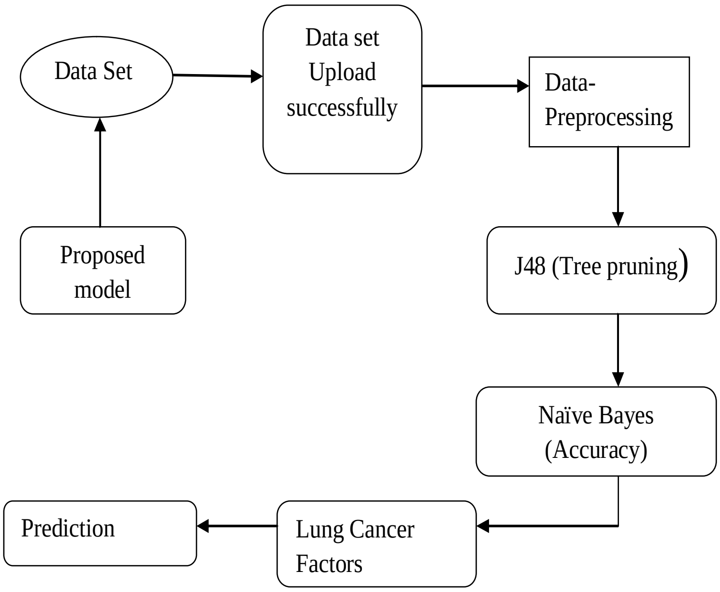
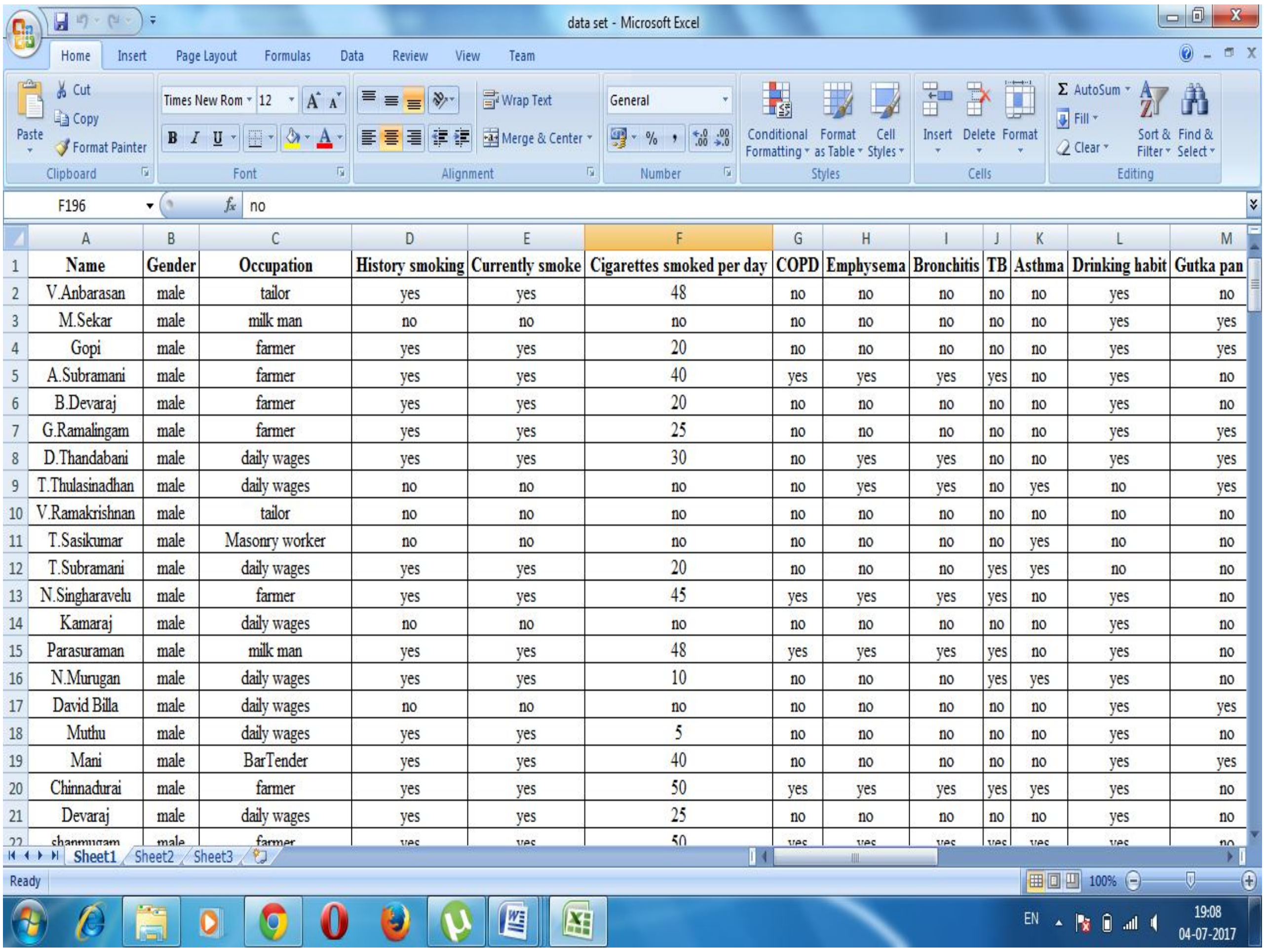


Figure 1 Proposed Architecture

4.1. DATA SET

The data collection for the study has been accomplished. These data were obtained from the various hospitals, comprising the lung cancer patients. The biggest risk factors that can be used to predict lung cancer, such as smoking history, cigarettes_smoking per day, emphysema,etc.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Gender	Occupation	History smoking	Currently smoke	Cigarettes smoked per day	COPD	Emphysema	Bronchitis	TB	Asthma	Drinking habit	Gutka pan
2	V.Anbarasan	male	tailor	yes	yes	48	no	no	no	no	no	yes	no
3	M.Sekar	male	milk man	no	no	no	no	no	no	no	no	yes	yes
4	Gopi	male	farmer	yes	yes	20	no	no	no	no	no	yes	yes
5	A.Subramani	male	farmer	yes	yes	40	yes	yes	yes	yes	no	yes	no
6	B.Devaraj	male	farmer	yes	yes	20	no	no	no	no	no	yes	no
7	G.Ramalingam	male	farmer	yes	yes	25	no	no	no	no	no	yes	yes
8	D.Thandabani	male	daily wages	yes	yes	30	no	yes	yes	no	no	yes	yes
9	T.Thulasinadhan	male	daily wages	no	no	no	no	yes	yes	no	yes	no	yes
10	V.Ramakrishnan	male	tailor	no	no	no	no	no	no	no	no	no	no
11	T.Sasikumar	male	Masonry worker	no	no	no	no	no	no	no	yes	no	no
12	T.Subramani	male	daily wages	yes	yes	20	no	no	no	yes	yes	no	no
13	N.Singharavelu	male	farmer	yes	yes	45	yes	yes	yes	yes	no	yes	no
14	Kamaraj	male	daily wages	no	no	no	no	no	no	no	no	yes	no
15	Parasuraman	male	milk man	yes	yes	48	yes	yes	yes	yes	no	yes	no
16	N.Murugan	male	daily wages	yes	yes	10	no	no	no	yes	yes	yes	no
17	David Billa	male	daily wages	no	no	no	no	no	no	no	no	yes	yes
18	Muthu	male	daily wages	yes	yes	5	no	no	no	no	no	yes	no
19	Mani	male	BarTender	yes	yes	40	no	no	no	no	no	yes	yes
20	Chinnadurai	male	farmer	yes	yes	50	yes	yes	yes	yes	yes	yes	no
21	Devaraj	male	daily wages	yes	yes	25	no	no	no	no	no	yes	no
22	shanmugam	male	farmer	yes	yes	50	yes	yes	yes	yes	yes	yes	no

Figure – 2 Sample data

4.2. TOOLS AND TECHNIQUES

4.2.1. DATA MINING CLASSIFICATION METHODS

Data mining, the extraction of hidden predictive information from large databases, Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

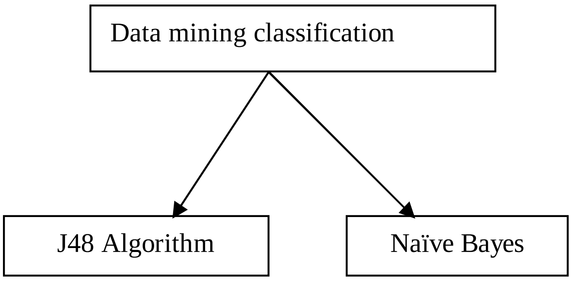
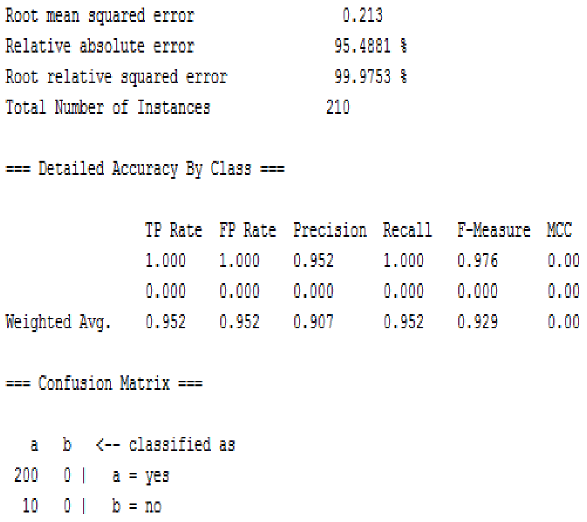


Figure 3 Data mining techniques in proposed work

4.2.1.1. J48

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly.

Performance Measure using J 48 Algorithm



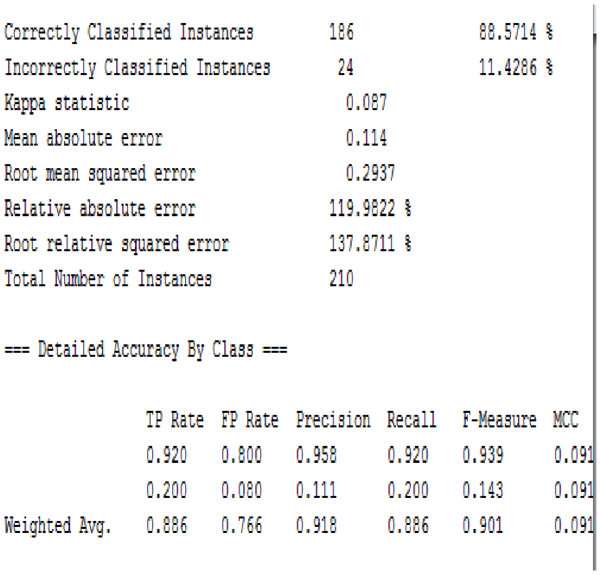
```
Root mean squared error                  0.213
Relative absolute error                 95.4881 %
Root relative squared error             99.9753 %
Total Number of Instances              210

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 1.000    1.000    0.952      1.000    0.976      0.00
                 0.000    0.000    0.000      0.000    0.000      0.00
Weighted Avg.    0.952    0.952    0.907      0.952    0.929      0.00

=== Confusion Matrix ===

   a   b   <-- classified as
 200   0 |   a = yes
  10   0 |   b = no
```

4.2.1.2 NAIVE BAYES

The naïve bayes model is a simple and well-known method for performing supervised learning of a classification problem. Assuming that the contribution by all attributes are independent that each contributors equally to the classification problem.

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

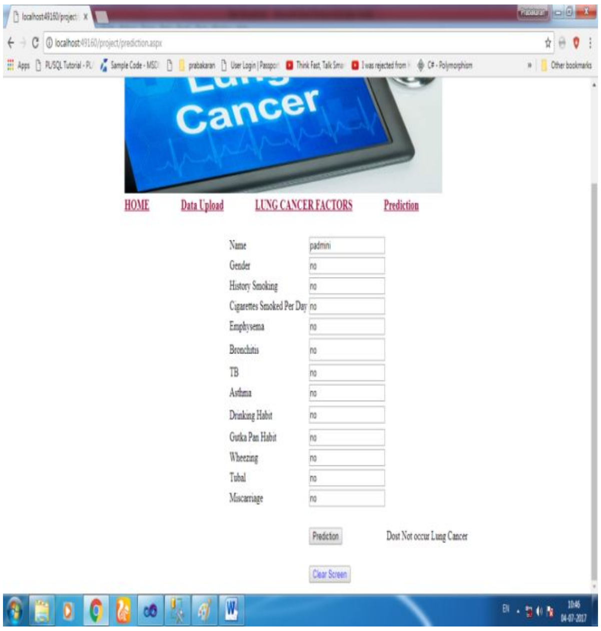
Performance Measure using Naive Bayes Algorithm



```
Correctly Classified Instances         186               88.5714 %
Incorrectly Classified Instances        24               11.4286 %
Kappa statistic                          0.087
Mean absolute error                      0.114
Root mean squared error                  0.2937
Relative absolute error                119.9822 %
Root relative squared error            137.8711 %
Total Number of Instances              210

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.920    0.800    0.958      0.920    0.939      0.091
                 0.200    0.080    0.111      0.200    0.143      0.091
Weighted Avg.    0.886    0.766    0.918      0.886    0.901      0.091
```

4.2.1.3. PREDICTION

The results thus obtained from the above classification technique is applied to the proposed component designed using ASP.NET which is capable of predicting the lung cancer effectively.



Name	padmini
Gender	no
History Smoking	no
Cigarettes Smoked Per Day	no
Emphysema	no
Bronchitis	no
TB	no
Asthma	no
Drinking Habit	no
Gutka Pan Habit	no
Wheezing	no
Tubal	no
Miscarriage	no

Dost Not occur Lung Cancer

5. RESULT AND DISCUSSION

The experiment has been performed using J48 algorithm and Naïve Bayes data mining classification techniques and it is found that the Naive Bayes algorithm gives a better performance over the other classification algorithm such as Bayesian and J48. LDPS (Lung Disease prediction system) produce more accuracy in the earlier stage by considering the factors used for prediction.

PERFORMANCE COMPARISON OF J-48 AND NAÏVE BAYES

Techniques/Me Assures	Correctly Classified Instances	Mis-Classified Instances	Overall Accuracy
J-48	200	10	95.23
Naïve Bayes	186	24	88.57

6. CONCLUSION AND FUTURE WORK

The proposed method is used for prediction of lung diseases in the earlier stage using data mining techniques by considering the factors which has high probability. This methodology is focussed on data collection based on questionnaire method from the common man. Hence it is limited to the knowledge of the individual. In future the lung diseases can be predicted in the earlier stage by considering the other data mining techniques which helps in identification of clinical prognostic factors, allowing individualization of patient's treatment as well as improved quality of anatomic imaging of the tumour and the regional lymph nodes, which results in a precise definition of the target volume.

REFERENCE

[1] Lawrence A. Loeb, Virginia L. Ernster, Kenneth E. Warner, John Abbotts, and John Laszlo "Smoking and Lung Cancer", on July 17, 2014.

[2] Kawsar Ahmed1, Abdullah-Al-Emran2*, Tasnuba Jesmin1,Roushney Fatima Mukti2, Md Zamilur Rahman1, Farzana Ahmed3, “Early Detection of Lung Cancer Risk Using Data Mining”, Asian Pacific Journal of Cancer Prevention, Vol 14, 2013.

[3] V.Krishnaiah, Dr.G.Narsimha, R.N.Subhash Chandra, “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques“(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013.

[4] Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh, “Mining Lung Cancer Data And Other Diseases Data using Data Mining Techniques: A Survey” Volume 4, Issue 2, March – April (2013).

[5] T.Karthikeyan, P.Thangaraju, “PCA-NB Algorithm to Enhance the Predictive Accuracy” International Journal of Engineering and Technology (IJET), Vol 6 No 1 Feb-Mar 2014.

[6] P.Thangaraju, G.Barkavi, “Lung Cancer Early Diagnosis Using Some Data Mining Classification Techniques: A Survey” COMPUSOFT, An international journal of advanced computer technology (IJACT), 3 (6), June-2014 (Volume-III, Issue-VI)

[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).