International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-2S2, January 2019

POPD Disease Diagnosing and Predictions Using Data Mining Algorithms

R.Jeena, P.Sarasu

Abstract: ***Data Mining is employed to seek out hidden pattern from a vast database. In data mining, machine learning is primarily act as research and from complex patterns which find decisions based on data. Persistent Obstructive Pulmonary Disease (POPD) is an important term accustomed describes progressive respiratory organ diseases as well as respiratory illness, bronchitis, and refractory (non-reversible) respiratory illness. This illness is characterized by increasing shortness of breath. These days Persistent Obstructive Pulmonary Disease (POPD) is one amongst the foremost causes of death within the developing countries. POPD is the main causes of lung cancer. By classification, common thoracic surgery includes information, procedural skill and decision to diagnose and treat diseases of the lungs. In this paper the data classification is Thoracic Surgery (Lung Cancer) patients' data set which includes 470 instances with 14 attributes is collected respectively. Data mining algorithms plays vital role in disease diagnosing and prediction. Among them we cannot access all DM algorithms. So, in this paper presents, chosen best one algorithm among the Standard DM algorithms for reduce the search and time complexity.***

Index Terms: ***Lung Tumor, Thoracic Surgery, Diagnosis, Naive Bayes, Random Forest, OneR, PART, Decision Stump, J48.***

I. INTRODUCTION

UCI Machine Learning Repository is a collection of different medical datasets [7]. POPD disease affects lungs which causes lung cancer. UCI Repository contains Thoracic surgery data. This data was composed from lung patients who experienced main lung surgery at Wroclaw Thoracic surgery Centre, connected with the department of Thoracic surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland. Thoracic surgery data contains 470 data on 17 several variables where 400 patients survived. This survey is about 70 patients didn't live a minimum of one year when the surgery. Average ages of the patients between 21 years and 87 years, in this maximum number of patients above 55 years of old. No missing data in the whole data set.

Manuscript published on 30 January 2019.
***** Correspondence Author (s)
Mrs.R.Jeena, Research Scholar, Department of CSE, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Tamilnadu, India.
Dr.P.Sarasu, Professor, Department of CSE, Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Tamilnadu, India.

II. RELATED WORK

In the work of Maciej Zieba et al. [1], used enhanced SVM for diagnosing post-operative anticipation. During this paper, they applied feature choice methodology to gauge the method. Since a current drawback with information sets is that little cardinality inscriptions the result, Shahian et al. [2] urged usage of artificial neural network to beat this drawback. Sindhu et al. [3] used numerous classification techniques to analyses body part surgery information and that they found that j48 provides higher accuracy. Harun et al [4] used completely different data processing algorithms and that they found Naïve Bayes is that the best.

III. RESEARCH METHODOLOGY

Model Diagram

Sequence of steps used in this research is summarized in Figure.1 below in the formulation of model diagram.

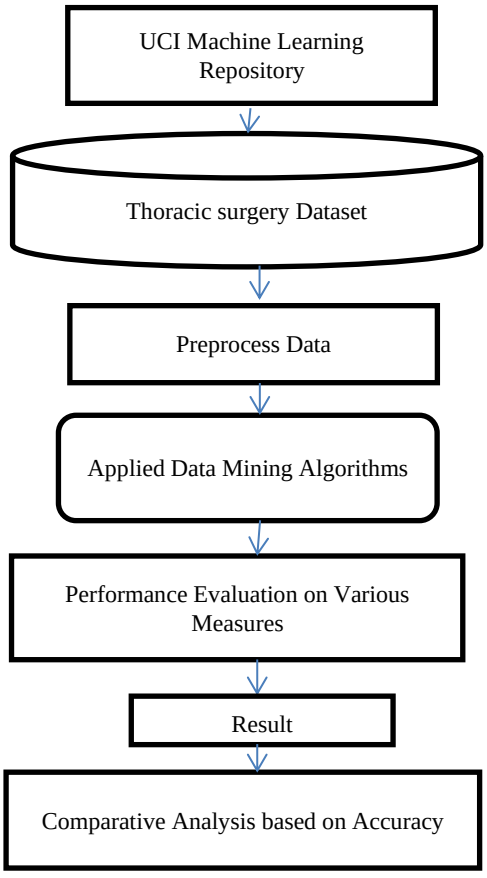

Fig.1. Proposed Model Diagram

B. Experimental setup

In this work, Weka tool is used for analysis the data. Weka is a software which is developed in the year 1997 by the students of Waikato University, New Zealand. Weka software is written Java programming language and uses GNU general public license (GPL). WEKA uses a Graphical User Interface to interact with the user and data files.

C. Algorithms Used

In this study, subsequent methods are used for prediction purpose.

1. Naive Bayesian Classifier

It is a probability method for resolving classification problems. Consider the variables A and C.

Joint probability: Pr(A=a,C=c)

Conditional probability: Pr(C=c | A=a)

The relationship between joint and conditional probability distributions Pr(C, A) = Pr(C | A) × Pr(A) = Pr(A|C) ×Pr(C)

Bayes Theorem:

P(C|A)=P(A|C)P(C) / P(A)

2. Support Vector Machine

Support Vector Machine (SVM) is a classification method for classifying both linear and nonlinear data. This method practices a nonlinear mapping to convert actual input data into a complex dimension. Within this novel dimension it discovers for the linear best separating hyperplane (that could be a "decision boundary" divides the records of 1 category from another).This hyperplane will be found by SVM using support vectors (which are the important training tuples) and margins. A separating hyperplane can be written as $\boldsymbol{W}$. $X+b = 0$, where $\boldsymbol{W}$ is a weight vector and b is a scalar, called as a bias.

3. Decision Tree Classifier

Decision tree is a supervised machine learning algorithm. It is used for classification problem. Decision tree is like a flow chart, which contains internal node, branch and leaf node. The top node is called as root node.

D. Cross-validation

For this study, standard of 10 fold cross-validations has been used. For this study, principle of 10 fold cross-validations has been used .In this method, at first; 10 equal sized data sets are produced from given knowledge. Then every knowledge set is divided into a pair of groups- 90% for coaching & ten anticipating testing. After that, a classifier is created with AN algorithmic program from ninetieth labeled knowledge and applied to one to 10 testing data for set 1.This procedure is sustained for set a pair of through ten. Within the final part, performance of the classifiers created from ten equal sized (training & testing) sets are averaged.

E. Performance criteria

After conducting a tenfold cross-validation of the info set, performances of the algorithms were analyzed by three metrics accuracy, F live and mythical monster curve. Accuracy decides the proportion of explanations that were properly categorized by the formula. Because it offers a basic presentation of every formula, accuracy was an honest start line of our analysis. Similarly, F live was a crucial applied mathematics analysis of classification because it measures take a look at accuracy. It's mean value of preciseness & recall. Finally, mythical monster curve was conjointly used as a good methodology of evaluating the quality or performance of expected copies here segment of true positives is aforethought beside the section of incorrect positives. House beneath the mythical creature arch is used for finding accuracy of models.

F. Hypothesis & test statistic

The aim of this learning is to match the presentations of varied processing techniques & their boosted versions for calculating survivability of pectoral surgery patients. For this learning, the hypothesis has been supported as monitors- Null hypothesis- all processing methods do similarly fine in forecasting conclusion of pectoral surgery different hypothesis boosted straight forward provision regression will the higher job. The datum utilized in testing this hypothesis, is modified paired t check as normal t- test will manufacture too several important variations due to dependencies within the estimates [6]. We tend to use this datum at ninety five confidence level.

G. Data Set Description

The data were collected retrospectively at the urban center body part Surgery Centre for patients United Nations agency have undergone major respiratory organ resections for primary carcinoma within the years 2007 to 2011. Body part Surgery (Lung Cancer) patients' information set is developed by assembling data from the hospital repository consists of 470 instances with fourteen totally different attributes. The Centre is connected to the Department of body part Surgery of the Medical University of urban focus and Lower Silesian Centre for pulmonic Diseases,Poland, whereas the analysis information establishes a component of the National carcinoma register, managed by the Institute of T.B. and pulmonic Diseases in capital of Poland, Poland. Table 1 shows the Thoracic Surgery data.

Table 1: Thoracic Surgery data

Attributes Name	Type	Attribute description
PRE4	Numeric	Forced vital capacity - FVC (numeric)
PRE5	True, False	A volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
PRE6	PRZ2, PRZ1, PRZ0	Performance status - Zubrod scale
PRE7	True, False	Pain before surgery
PRE8	True, False	Haemoptysis before surgery
PRE9	True, False	Dyspnea before surgery
PRE10	True, False	Cough, before surgery
PRE11	True, False	Weakness before surgery
PRE14	OC11, OC14,	T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest)
PRE17	OC12, OC13	Type 2 DM - diabetes mellitus
PRE19	True, False	MI up to 6 months
PRE25	True, False	PAD - peripheral arterial diseases
PRE30	True, False	Smoking
PRE32	True, False	Asthma

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

AGE	Numeric	Age at surgery
RISK1Y	True, False	1 year survival period - (T)rue value if dead

IV. EXPERIMENTAL RESULTS

A. Data Preparation

The variables are already classified and portrayed by numbers. The way within which the collision occurred has been classified as 3.

1. Based on Diagnosis

Diagnosis is that the elaborated combination of ICD10 codes for primary and secondary similarly as many tumour if some (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1).Table-2 shows the analysis results for the higher than 3 strategies reveals that the performance victimization the researchers planned prediction to enhance the performance of all the 3 classification methods, and therefore the noticeable improvement is that logistical regression produces a really important amendment in its performance represents one hundred and ninety five.65 that outperforms the opposite classification strategies. The feature ranking techniques are supported 3 completely different subsets of attributes like identification, Performance and tumour size. The remaining fourteen attributes set consists of performance of the tumour patients' details[8]. From the results obtained it's ascertained that the attribute with the 3 completely different classifier algorithmic rules have the Random Forest classifier is that the best algorithm and better accuracy than the opposite set of algorithms. Therefore herewith it's over that the 3 main attributes with 3 algorithms, the random forest is that the future process rather than victimization the opposite attributes.

2. Based on Performance

Performance status - Zubrod scale (PRZ2, PRZ1, and PRZ0). There are various ways of assessing general health. the globe Health Organization designed the size that doctors use most frequently. It has classes from 0 to 5.

0 – you are completely energetic and a lot of or less as you were before your sickness.

1- You cannot perform significant physical work, however will do anything.

2–You are up and concerning over 1/2 the day and may take care of yourself, however don't seem to be to an adequate degree to figure

3 – You are in bed or sitting during a chair for over 1/2 the day and you would like some facilitate in taking care of yourself

4– You are sleeping or a seat constantly and need heaps of dealing with yourself.

5 – Death

3. Based on Size of the tumor

T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13).

B. Experiment Steps in Weka

Step:1 - In 'Preprocess' tab click on 'Open URL' button and paste the link "https://archive.ics.uci.edu/ml/ machine-learning-databases/00277/ThoracicSurgery.arff" file. Figure 1 shows the preprocess window.

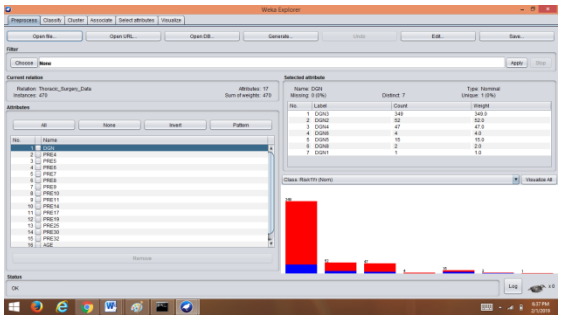

Fig. 1 : Preprocess window

Step:2 - Next select the "classify" tab and click choose button to select the "NavieBayes" in the classifier. Figure 2 shows the run information of Naïve Bayes classifier and Figure 3 shows the run information of SVM.

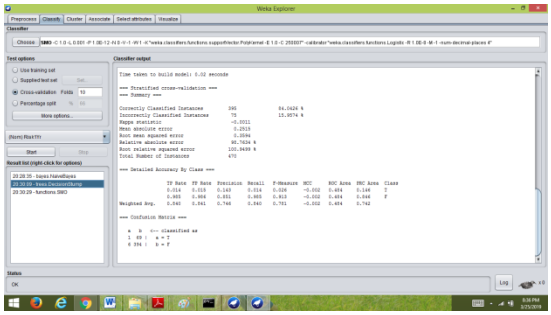

Fig. 2 : Naïve Bayes execution report

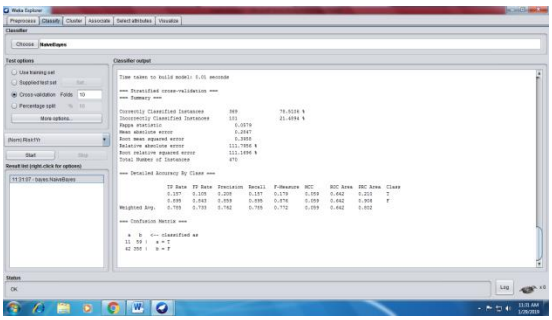

Fig. 3 : SVM Execution report

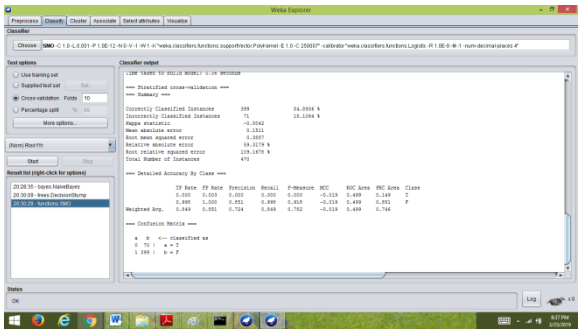

Fig. 4 : Decision Tree Execution report

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

C. Comparative Analysis

Table 2 shows the different performance values of the three data mining algorithms such as Naïve Bayes Classifier, Support Vector Machine and decision Tree.

Table: 2 – Analysis results

Attributes	Naïve Bayes	SVM	J48
Accuracy	91.91	82.34	85.1
Precision	0.763	0.718	0.711
Recall	0.823	0.766	0.837
F-Measure	0.782	0.74	0.769

Figure 5, Figure 6, Figure 7 and Figure 8 shows the comparison chart of difference performance measures.

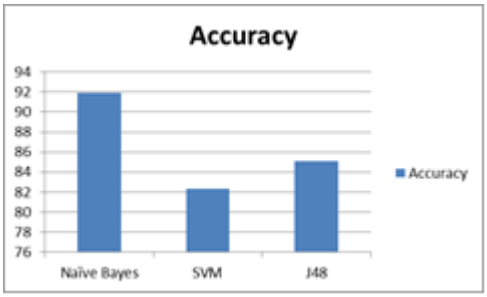

Fig. 5: Accuracy Measure

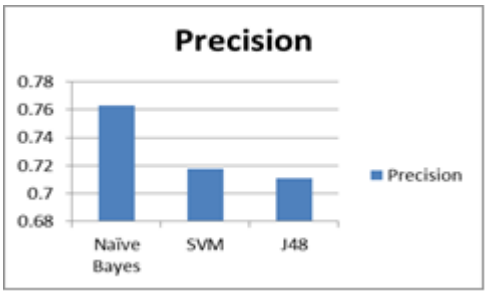

Fig. 6: Precision Measure

Figure 7: Recall Measure

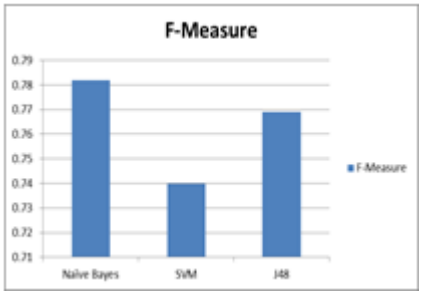

Fig. 8: F- Measure

V. CONCLUSION

The aim of this paper is to sight the causes of respiratory lung cancer. The dataset for the study contains thoracic Surgery and its wholly seventeen attributes of the year 2013 made by the Department of body part Surgery of the Medical University of metropolis and Lower-Silesian Centre for pneumonic Diseases, Republic of Poland and investigates the performance of Naive Bayes, SVM and Decision Tree for predicting classification accuracy. The classification accuracy of the check result reveals the subsequent 3 cases like diagnosing, Performance and tumour size. Once victimization the Naïve Bayes algorithmic rule[9], it shows the properly classified proportion values for all varieties. Naïve Bayes classifiers offer additional accuracy than a SVM classification algorithmic rule. The rule induction generates solely correct rules supported the accuracy. The choice Stump offers the read (values and classes). During this analysis, it's been found that the Naïve Bayes is that the best within the thoracic surgery for predicting the survival after 1 year of thoracic surgery.

REFERENCES

1 M. Zięba, J .M. Tomczak, M. Lubicz, and J. Świątek, “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients” , Applied Soft Computing, vol. 14, pp 99-108, Jan. 2014.

2 D. M. Shahian, S. L. Norman, D.F. Torchiana, “Cardiac Surgery Report Cards: Comprehensive Review and Statistical Critique” , Annals of Thoracic Surgery, vol. 72, pp 2155-2168, 2001.

3 V. Sindhu, S. A. S. Prabha, S. Veni , and M. Hemalatha, “Thoracic surgery analysis using data mining techniques” , International Journal of Computer Technology & Applications , vol. 5 pp 578-586, May,2014 .

4 Md. Ahasan Uddin Harun, Md. Nure Alam “Predicting Outcome of Thoracic Surgery by Data Mining Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 1, January 2015.

5 R. E. Schapire, (1990). "The Strength of Weak Learnability" Boston, USA: Kluwer Academic Publishers, 1990.

6 J. R. Quinlan, C4.5: Programs for Machine Learning, 1st edition, Massachusetts, USA, Morgan Kaufmann , 1993.

7 (2014) The UCI Machine Learning website. [Online]. Available: http://www.archive.ics.uci.edu/ml/ [11] S. Haykin , Neural Networks : a comprehensive foundation, 1st edition, London, Prentice Hall, 1999.

8 T. Dietterich , “approximate statistical tests for comparing supervised classification learning algorithms” , Neural Computation, vol. 10, pp1895-1924, 1998.

9 V. K. Mago and N. Bhatia, Cross-disciplinary Applications of Artificial Intelligence and Pattern Recognition : Advancing Technologies, 1st ed., Pennsylvania, USA : IGI Global, 2011.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)