Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 7, July 2019, pg.88 – 97

PERFORMANCE COMPARISION OF DATA CLASSIFICATION ALGORITHM FOR ANALYSIS LUNG CANCER DATASETS

Dr. M.Mayilvaganan[1]; N.Thamaraikannan[2]

Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[1]

Email Id – mayil24_02@yahoo.co.in

Research Scholar, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[2]

Email Id – n.thamaraikannan57@gmail.com

ABSTRACT: Lung cancer is the major cause of cancer deaths in both men and women in world wide. Lung malignancy begins when cells of the lung begin unusual cells grow into the lungs. Various types of approaches have been used for lung cancer diagnosis. There are two kinds of lung cancer that is small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC).People who smoke have greatest risk of carcinoma. The risk of carcinoma will increase with the length of your time and variety of cigarettes they need preserved. To gather the information and to measure the data's based on smokers, Nonsmokers in keeping with their age and history of habit. In this research works mainly specialized in lung cancer data and it uses machine learning techniques Naive bayes and decision tree. The main objective of this technique used for classification and prediction and mainly focused on performance comparison of machine learning algorithm with respective run time execution of each algorithm. The naïve bayes and decision tree algorithm used to analysis for performance comparison accuracy of each algorithm. We were using the Rapid Miner software to predict the accuracy of Classification algorithms. In this comparison we use Lift Chart and ROC curve in which it displays the accurate value of classification.

INTRODUCTION

Data mining refers to extracting information from vast volume of data.its commonly used in proflie practices such as marketing, survalience, frauddection, customer retentation, production control, science exploration.

Data mining have been drawing in a lot of consideration in the data business and in the public eye in general. Because of accessibility of wide information an unavoidable requirement for preparing those information into learning. Those data or information picked up can be utilized for applications ranging from market examination, misrepresentation discovery, and client maintenance, to create control and science investigation.

DATA MINING IN KDD

Data Mining, additionally famously known as Knowledge Discovery in Databases (KDD), alludes to the nontrivial extraction of verifiable, beforehand obscure and possibly valuable data from information in databases. While information mining and learning revelation in databases (or KDD) are as often as possible regarded as equivalent words, information mining is entirely of the learning disclosure process.

Data mining is only one of the many steps involved in knowledge discovery in databases.the various steps in the knowledge discovery process include data selection,data cleaning and preprocessing,data transformation and reduction,data mining algorithm selection and finally the post processing and the interpretation of the discovery knowledge.The KDD process is highly iterative and interactive.

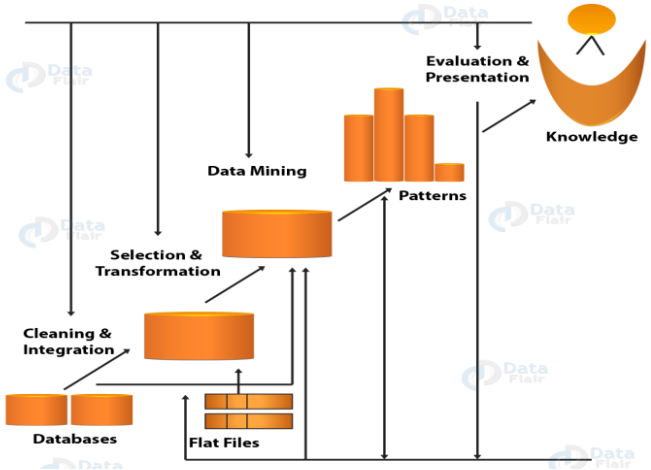


Fig 1.1 Knowledge Discovery in Databases Process

1.Data Cleaning

In this phase as cleaning data and remove noise data and unrelated data.

2.Data Integration

In this phase multiple data source may be integrated in common source.

3.Data Selection

In this phase,the data applicable to the analysis is decided on recover from the data.

4.Data Transformation

In this phase transforned in order to be acceptable task of data.

5.Data Mining

In this phase extraction of patterns from data.

6.Pattern Evalution

In this phase strictly interesting pattern representing information are identified based on given measured.

7. Knowledge Representation

In this final phase in which the find the information is visually represented to the user. This requirment step uses visualization techniques to understand and explain the data mining results.

LUNG CANCER

Lung disease is a growth that begins in the lungs. Lung disease is the most widely recognized reason for malignancy demise around the world. The event of lung growth has expanded quickly and turned into the most widely recognized malignancy in men in many nations. Smoking is by a wide margin the most critical preventable reason for disease on the planet. In the event that the first lung disease has spread, a man may feel side effects in different places in the body. Normal spots for lung malignancy to spread incorporate different parts of the lungs in Tumor, lymph hubs, bones, cerebrum, liver. The frequency of lung tumor is unequivocally connected with cigarette smoking, with around 90% of lung growths emerging because of tobacco utilize. The danger of lung disease increments with the quantity of cigarettes smoked after some time. The danger of lung disease increments with the quantity of cigarettes smoked after some time.

There are two type of lung cancer

- **Small Cell Lung Cancer(SCLC)**

  Its construct the 15-20% of lung cancer.SCLC tends to growth in the middle of lungs and regularly spread more immediate than NSCLC

- **Non Small Cell Lung Cancer(NSCLC)**

  Its construct 80% of lung cancer.NSCLC classifief by

  - Adeno carcinoma
  - Squamous cell carcinoma
  - Large cell carcinoma

CAUSE OF LUNG CANCER

- Smoking
- Passive smoking
- Radon
- Asbestos
- Air pollution

SYMPTOMOS

- Cough
- Breathlessness
- Chest pain
- Wheezing
- Fatigue and weakness
- Clubbing of the fingernails
- Fever

- Shortness of breath
- Loss of appetite

SCOPE OF THE RESEARCH

The scope of the research is to accuracy of data. This system compares the classification of naive bayes and decision tree machine algorithm to find which is better. We classify the data based on accuracy, Precision, Accuracy and Recall. We finally predict the data and remove the sensitive data.

NAÏVE BAYES ALGORITHM

The Bayesian Classification speaks to an administered learning strategy and in addition a measurable technique for characterization. Accept a fundamental probabilistic model and it enables us to catch vulnerability about the model principally by deciding probabilities of the results. It can take care of analytic and prescient issues. This Arrangement is named after Thomas Bayes (1702-1761), who proposed the Bayes Hypothesis.

Bayesian arrangement gives commonsense learning calculations and earlier information and watched information can be consolidated. Bayesian Classification gives a valuable viewpoint to comprehension and assessing many learning calculations. It figures express probabilities for theory and it is hearty to commotion in information.

The overall precision for a classifier for a given dataset, average of precision of both classes is calculated. Bayes hypothesis gives a method for ascertaining the back likelihood, P (c|x), from P(c), P(x), and P (x|c). Credulous Bayes classifier thinks about that the impact of the estimation of an indicator (x) on a given class (c) is free of the estimations of different indicators.

Problem definition

- A training set X, where each training instance x is represented as an n-dimensional attribute vector: (x1, x2…x_n)

- A pre-defined set of classes: C= {c1, c2... c_m}

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

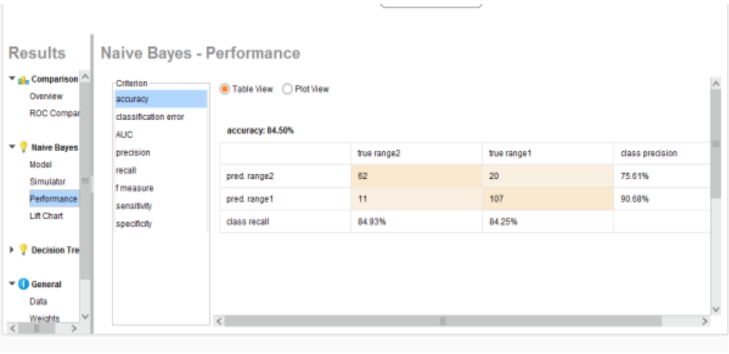
$$P(c|X) = P(x_1|c) * P(x_2|c) * \ldots\ldots\ldots. *P(x_n|c)*P(c)$$

P(c|x) is the posterior probability of class given predictor od class.

P(c) is called the prior probability of class

P(x|c) is the likelihood which is the probability of predictor of given class

P(x) is the prior probability of predictor of class.



Naive Bayes - Performance

accuracy: 84.50%

	true range2	true range1	class precision
pred. range2	62	20	75.61%
pred. range1	11	107	90.68%
class recall	84.93%	84.25%	

Fig 1.2 PERFORMANCE OF NAÏVE BAYES ALGORITHM

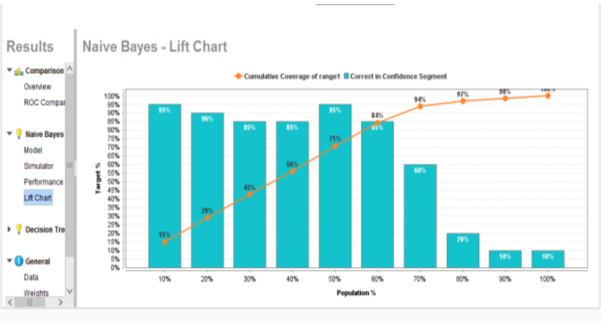


Fig 1.3 LIFE CHART AND ROC CURVE OF NAÏVE BAYES ALGORITHM

DECISION TREE

The decision tree is a work of art and common model of learning. It is firmly identified with the essential software engineering thought of "divide and prevail." Although decision trees can be used too many learning problems, we will start with the simplest case: binary classification.

Classification tree method use to classify data. the method gas establish and use to various environment medical, market research,statisticalanalysis,masketking and customer relationship. The various types of application use classification tools in huge data's. The major use of decision tree to uncover the formation that contained into data and arrange the data.

- Decision tree can hold both nominal and numeric input values.
- Decision tree are read to non-parametric method.
- Decision tree are fit for taking care of datasets that may have missing qualities.
- Decision tree are efficient to holding datasets that have missing values

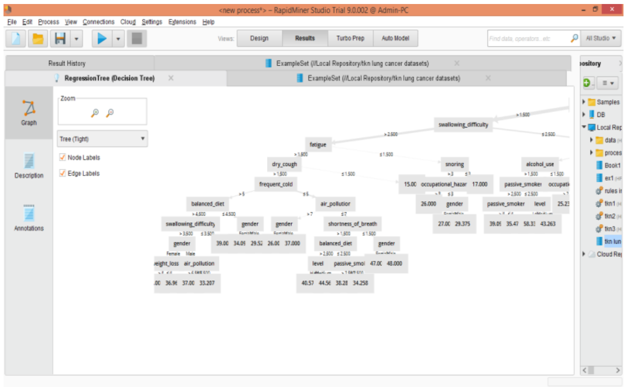
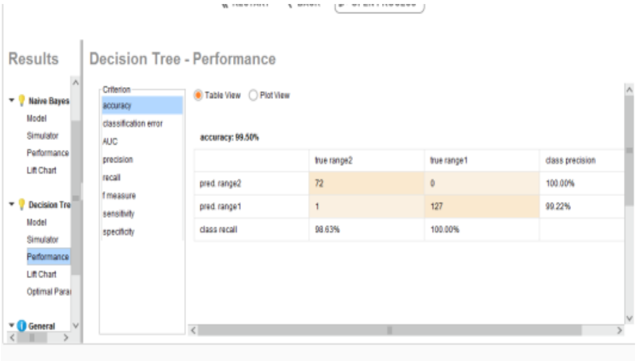


Fig 1.4 FORMATION OF DECISION TREE



Decision Tree - Performance

accuracy: 99.50%

	true range2	true range1	class precision
pred. range2	72	0	100.00%
pred. range1	1	127	99.22%
class recall	98.63%	100.00%	

Fig 1.5 PERFORMANCE OF DECISION TREE ALGORITHM

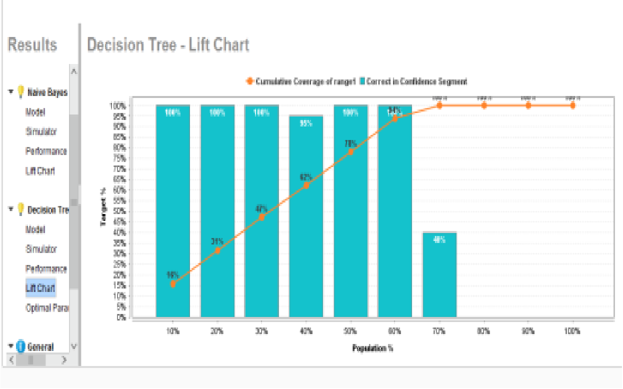
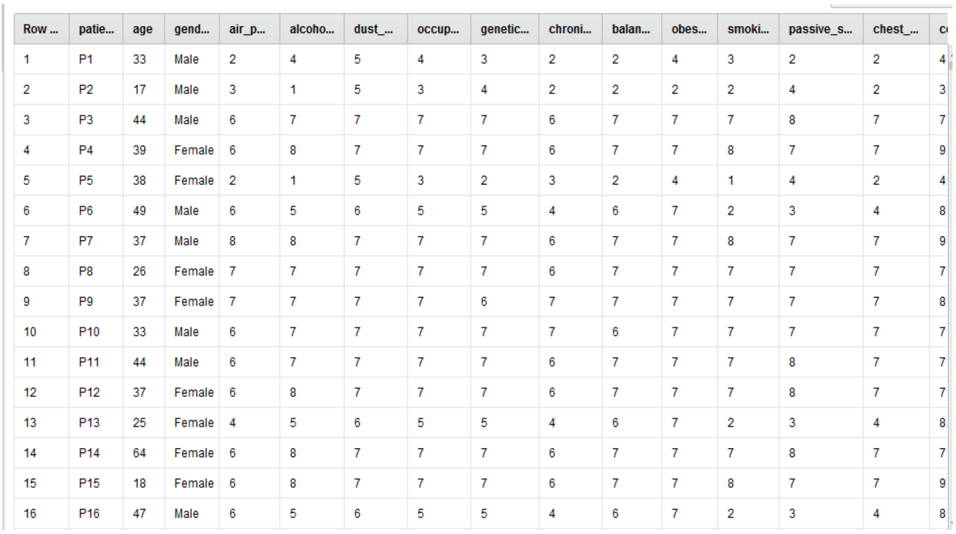


Fig 1.6 LIFE CHART AND ROC CURVE OF NAÏVE BAYES ALGORITHM

DATA FOR RESEARCH

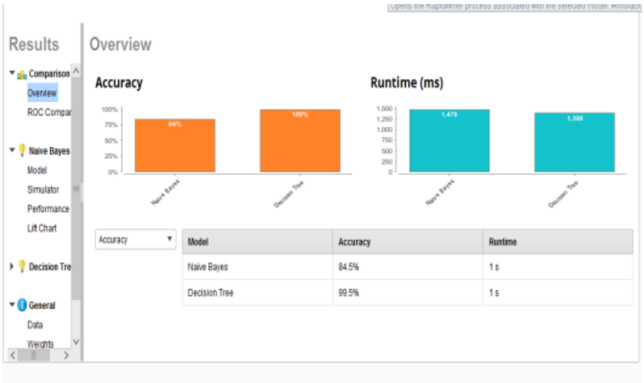


Row ...	patie...	age	gend...	air_p...	alcoho...	dust_...	occup...	genetic...	chroni...	balan...	obes...	smoki...	passive_s...	chest_...	c
1	P1	33	Male	2	4	5	4	3	2	2	4	3	2	2	4
2	P2	17	Male	3	1	5	3	4	2	2	2	2	4	2	3
3	P3	44	Male	6	7	7	7	7	6	7	7	7	8	7	7
4	P4	39	Female	6	8	7	7	7	6	7	7	8	7	7	9
5	P5	38	Female	2	1	5	3	2	3	2	4	1	4	2	4
6	P6	49	Male	6	5	6	5	5	4	6	7	2	3	4	8
7	P7	37	Male	8	8	7	7	7	6	7	7	8	7	7	9
8	P8	26	Female	7	7	7	7	7	6	7	7	7	7	7	7
9	P9	37	Female	7	7	7	7	6	7	7	7	7	7	7	8
10	P10	33	Male	6	7	7	7	7	7	6	7	7	7	7	7
11	P11	44	Male	6	7	7	7	7	6	7	7	7	8	7	7
12	P12	37	Female	6	8	7	7	7	6	7	7	7	8	7	7
13	P13	25	Female	4	5	6	5	5	4	6	7	2	3	4	8
14	P14	64	Female	6	8	7	7	7	6	7	7	7	8	7	7
15	P15	18	Female	6	8	7	7	7	6	7	7	8	7	7	9
16	P16	47	Male	6	5	6	5	5	4	6	7	2	3	4	8

RESULTS AND DISCUSSION

The comparison of Machine Learning algorithm Naïve Bayes and Decision Tree algorithm using Rapid Miner studio. We have classifier analyzed large volume of data. The Naïve Bayes classifiers provides the highest accuracy of 84.5% per 1 Second and Decision tree classifier provides 99.5% per 1 Second accuracy provided. In this research work, we have found this one is the effective algorithm to identify their highest performance in runtime.

PERFROMANCE COMPARISON FOR NAÏVE BAYES



Model	Accuracy	Runtime
Naive Bayes	84.5%	1 s
Decision Tree	99.5%	1 s

Fig 1.7 ALGORITHM AND DECISION TREE ALGORITHM

CONCLUSION

In this research work was comparison between two various types of machine Learning classification algorithms based on the Lung cancer dataset. This comparison on classification algorithm helps to detect the accuracy of data. We used the Lift chart and ROC Curve to display the output for this comparison technique. So we can easily understand the results on lung cancer data. The performance of algorithm on this dataset have been evaluated and classified based on the Bayes Classifier. Therefore the Decision Tree algorithms performance is high and results the more accurate data set which is compared to decision tree. We came to a conclusion that Decision Tree algorithm is the best.

FUTURE WORK

This research work mainly deals with machine learning algorithm for early find of Lung Cancer. These techniques are useful for the detect of lung cancer then use to new classification algorithms which one can provide better solution. Further this research can be enhancing using big data, IoT, Deep Learning methods to increase the delicate and specificity and to provide accurate result.

BIBLIOGRAPHY

[1] Jiawei Han, MichelineKamber "Data Mining Concepts and Techniques" in proceeding of second edition Morgan Kaufmann Publisher an imprint of Elsevier 2006.

[2] AlkaGangrade, Ravindra Patel " SMC Protocol for Naïve Bayes classification over Grid Partitioned Data using Multiple UTPs" International Journal of Computer Applications(0975 – 8887) Volume 64- No 6. February 2013.

[3] Li Liu, Murat Kantarcioglu and BhavaniThurasingham"A Novel Privacy Preserving Decision Tree Algorithm" Technical Report October 2006.

[4] Ashmeet Singh, R Sathyaraj "A Comparison Between Classification Algorithms on Different Datasets Methodologies using Rapid miner" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 5, May 2016.

[5] JosipMesarić, Dario Šebalj"Decision trees for predicting the academic success of Students" Croatian Operational Research Review CRORR 7(2016), 367–388 December 30, 2016.

[6] ShahrukhTeli, PrashastiKanikar "A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[7] SunpreetKaur, Sonikaa Jindal," A Surveyon Machine Learning Algorithms", International Journal of Innovative Research in Advanced Engineering (IJIRAE),ISSN: 2349-2763,November 2016.

[8] KajareeDas,Rabi Narayan Behera," A Survey on Machine Learning: Concept, Algorithms and Applications", International Journal of Innovative Research in Computer and Communication Engineering,ISSN:2320-9801,February 2017.

[9] FatmaTaher, NaoufelWerghi and Hussain Al-Ahmad (2012), "Bayesian Classification and Artificial Neural Network Methods for Lung Cancer Early Diagnosis", IEEE.

[10] Thangaraju P, Karthikeyan T, Barkavi G, Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[11] Krishnaiah V, Narsimha G, Subhash Chandra N , Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45.

[12] ParagDeoskar, 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).

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

[14] National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N Engl J Med 2011; 365:395-409.