International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206669

Disease Prediction by Machine Learning Over Big Data Lung Cancer

T. Shanmuga Priya*, Dr. T. Meyyappan
Department of Computer Science, Alagappa University, Karaikudi, Tamil Nadu, India

Article Info
Volume 7, Issue 1
Page Number: 16-24
Publication Issue :
January-February-2021

Article History
Accepted : 01 Jan 2021
Published : 04 Jan 2021

ABSTRACT

Lung Cancer is one of the deadly diseases in the world today. Lung Cancer is caused because of some genetic factors and/or environmental factors and/or today's modern lifestyle. Lung cancer has become the primary reason of death in developed countries. The majority effective way to decrease lung cancer death is to detect it earlier. The in advance detection of cancer is not easier method but if it is detecte it is curable. Various works have been done in predicting lung cancer different data mining approach and algorithm were adopt by different people. All work has some limits such as lack of intelligent prediction, and incompetent in structure that forced to take up this problem and to implement the Data mining based cancer prediction System (DMBCPS). This has proposed the Lung cancer prediction system based on data mining. This system is validated by comparing its predicted results with patient's prior medical information and it was analyzed by using weka tool system. We analyzed the lung cancer prediction using classification algorithm such as Naive Bayes, SVM and Random forest algorithm. The dataset have 782 instances and 31 attributes. The main aim of this paper is to provide the earlier warning to the users and the performance analysis of the classification algorithms.

Keywords: Data mining, Lung Cancer Naive Bayes (NB), Support vector Machine (SVM), Random forest

I. INTRODUCTION

Data Mining is a knowledge mining process. It is interdisciplinary subfields of computer science. The tremendous growth of scientific databases put a lot of challenges before the research to extract useful information from them using traditional data base techniques. Hence effective mining technique is significant to discover the implicit information from huge databases. Cluster analysis is one of the major data mining algorithms, extensively used for a lot of practical application in various emerging areas like Bioinformatics. Clustering is an unverified method that subdivides an input data set into a desired number of subgroups so that the objects of the same subgroup will be similar or associated to one another and different from or unrelated to the objects in other groups. A high-quality clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The excellence of a clustering result depends on both the similarity measure used by the method and its execution and also by its ability to discover Some or all of the hidden patterns decision tree is a usually used clustering method that tries to find a user specified number of clusters parents, which are represent by their centroids, by minimizing the square error function. The clustering method aims at optimizing the cost purpose to minimize the difference of the objects within each cluster, while maximizing the dissimilarity of different clusters.

Lung cancer is the one of the most important cause of cancer death in both men and women. Symptom of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. Treatment and forecast depend on the histological type of lung cancer, the stage and the patient's performance status. Possible treatments include surgery, chemotherapy, and radiotherapy Survival depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer survive five existence after the diagnosis. Symptom that may suggest lung cancer include:

- dyspnea (shortness of breath with activity),
- hemoptysis (coughing up blood),
- chronic coughing or change in regular coughing pattern,
- wheezing,
- chest pain or pain in the abdomen,
- cachexia (weight loss, fatigue, and loss of appetite),
- dysphonia (hoarse voice),
- clubbing of the fingernails(uncommon),
- dysphasia(difficulty swallowing),
- Pain in shoulder ,chest , arm,
- Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

II. LETERATURE SURVEY

Agustin Blas et al.[8] described the performance of the grouping Genetic algorithm in clustering, started with proposed encoding, and different modification of crossover and mutation operation and also initiated the local search include with the island model for improve the performance of the problem. The real data sets like iris and wine were used and compared the results with the classical approaches such as DBSCAN and K-means, and obtaining the excellent results in proposed grouping based methodology the evolutionary approach such as Genetic algorithm. The performance of the algorithm was measured by using the different fitness function.

Tzung-Pei-Hong et al.[9] discussed the performance of the Genetic algorithm based attribute clustering process were improved based on the grouping Genetic algorithm. The chromosome representation, Genetic operations, and fitness function defined in grouping Genetic algorithm for solving the clustering problem. The result of grouping Genetic algorithm based clustering algorithm improved the convergence speed and fitness value of the clustering problem. In addition the algorithm can also deal with the problem of missing values. The other optimization algorithms are used to solve the problem in attribute grouping.

Daniel Gomes Ferrari et al. [10] proposed a new approach to characterize the clustering problems based on the similarity among objects and the method for combine internal indices for ranking algorithms based on the performance of the problem. The experimental results indicated the viability of meta learning systems for an unlabeled approach to the clustering algorithm selection problem. This technique presents the better result from the distance based set over the attribute based approach.

Kunnuri Lahari et al. [13] enhanced reduce the local minima using evolutionary and population based methods like Genetic algorithm and teaching learning based optimization. The data sets iris and wine are used, and the experimental results are compared with the Genetic algorithm and teaching learning based optimization based clustering with k-means algorithm. The performance of the evolutionary based clustering method compared with some existing clustering method.

Rahila H.Sheikh et al.[14] proclaimed a brief study of Genetic algorithm based clustering. Rajashree Dash et al.[11] discussed on comparative analysis of K-means and Genetic algorithm based on clustering. Arun Prabha et al.[15] with respect to the idea were improved the cluster quality from K-means clustering using a Genetic algorithm. Large scale clustering problems in data mining also address by this method. The best results are achieved by using this method.

Anusha et al.[16] depicted an enhanced K-means Genetic algorithm for optimal clustering. The author overcomes the disadvantage of local optima with suitable dataset and also the algorithm fails in computational time. It is inferred that the technique produced more than the 90% accuracy for real life dataset. The author also adopted a neighborhood knowledge strategy for optimizing multi objective troubles. This algorithm used k means Genetic algorithm to find the smallness of the clusters. It is noted that the algorithm could produce minimum index value for the maximum datasets.

III.METHODOLOGY

3.1 OVERVIEW OF DATAMINING

Data Mining is the innovation of hidden information found in large quantities of data and can be viewed as a step in the knowledge discovery process (Fayyad 1996).Data mining defined as a set of computer-assisted techniques designed to automatically mine big volumes of integrated data for new, hidden or unexpected information, or interesting patterns. With small set of data, traditional statistical analysis can be efficiently used. The first and simplest analytical step in data mining is to explain the data summarize its statistical attributes such as means and standard deviations, visually evaluation it using chart and graphs, and look for potentially meaningful links among variables such as values that often occur together (Edelstein 1998).

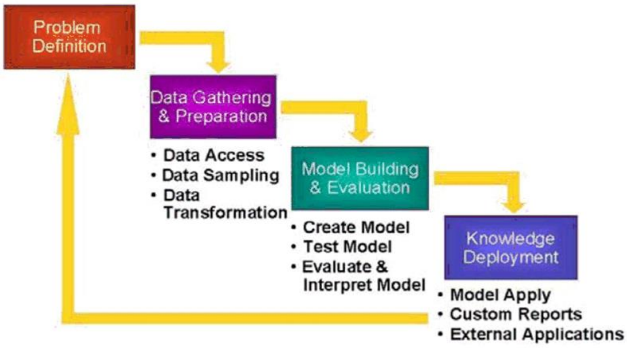


Figure 3.1 An overview of steps that compose KDD process

3.4 PROPOSED ALGORITHMS

3.4.1 SVM

SVM, a decently brand new sort of learning calculation, initially presented. Actually, SVM go for hyper plane incredible isolates the classes of information. SVMs has affirmed the ability not just too precisely isolate elements into right classes, additionally to recognize case whose build up arrangement is not upheld by information. In spite of the fact that SVM are nearly harsh characterize dispersion of preparing cases of every class. It is just reached out numerical figuring's.Two such expansion, the first is to make bigger out SVM relapse examination, where the objective to deliver a direct ability that can authentically exact that objective capacity. An further growth is to figure out how to rank components as opposed to creating a characterization for individual components. Positioning can be decreased to looking at sets of case and delivering a +1 assess if the combine is in the right positioning request not withstanding –1 generally.

Algorithm for SVM:

```
Step1: Select candidate = {closest pair of opp class}
Step2: while there are violating points do
Step3: Find a violator
Step4: Candidate = Candidate U violator
Step5: if any z < 0 due to addition of c to s then
Step6: Candidate = candidate/p
Step7: repeat till all such points are pruned
Step8: End of if
Step9: End of while
```

3.4.2 NAIVE BAYES

The Naïve Bayes Algorithm is a probabilistic calculation that is successive in nature, taking after ventures of execution, grouping, estimation and expectation. For discovering relations between the ailments, side effects and drugs, there are different information mining existing arrangement, however these calculations have their own confinements; various mphases, binning of the ceaseless contentions, high computational time, and so forth. Innocent Bayes conquers different restrictions including oversight of complex iterative estimations of the parameter and can be connected on a substantial dataset continuously.

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. This assumption is called class conditional independence.

$$P(C \mid F_1 .....F_n) = \frac{P(F_1 .....F_n \mid C) * P(C)}{P(F_1 .....F_n)}$$

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

P(c) is the prior probability of class.

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

P(x) is the prior probability of predictor.

3.4.3 Random Forest

Random forests are an together learning method for classification, regression and other tasks that operate by construct a multitude of decision trees at training execution time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests right for decision trees habit of over fitting to their training set.

Algorithm

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $A=A_1……,A_n$ with responses $B=B_1……,B_n$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

1. Sample, with replacement, n training examples from A, B; call these A_b, B_b.
2. Train a classification or regression tree f_b on A_b, B_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

or by taking the majority vote in the case of classification trees.

3.4.4 WEKA TOOL

Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time. The Weka or woodhen is an endemic bird of New Zealand. It provide many special algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Advantages

- As weka is fully implemented in java programming languages, it is platform independent & portable.
- It is freely available under GNU General Public License.
- Weka s/w contain very graphical user interface, so the system is very easy to access.
- There is very large collection of different data mining algorithms.

Disadvantages

- Lack of possibilities to interface with other software
- Performance is often sacrificed in favor of portability, design transparency, etc.

Memory limitation, because the data has to be loaded into main memory completely

IV.EXPERIMENTAL RESULT

The proposed system has been implemented using WEKA tool. The results are obtain as follow after execution. Sample data set is presented. Data a Lung cancer has been used with 782 items each for analysis. In this set of association rules are obtained by applying SVM, NB, RF algorithm. By analyzing the data, and giving different support and confidence values, execution time, can obtain different number of rules. During analysis it found that SVM is much faster for huge number of transactions as compare to NB & RF. It takes fewer time to generate frequent item sets.

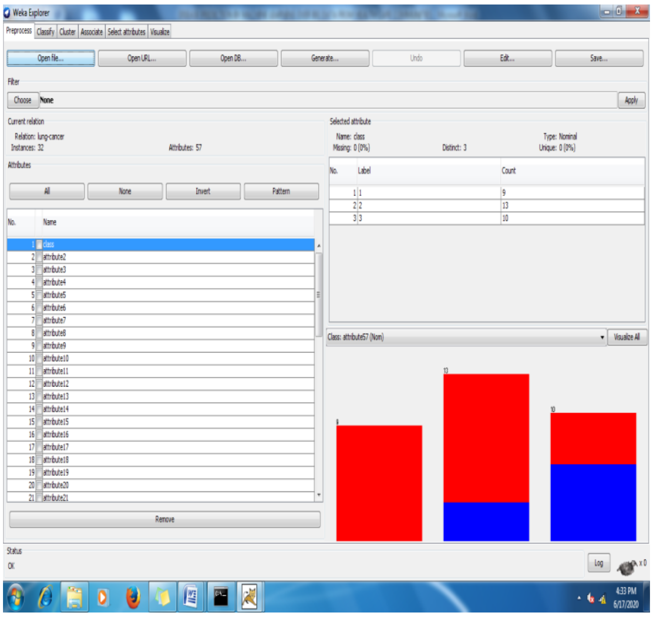
Dataset Description

Dataset used in this study is more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Attributes for symptom is used to diagnosis of disease are to be handled efficiently to obtain the optimal outcome from the data mining process. The attribute such as, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, passive smoker, chest pain, coughing of blood, Fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails, Frequent Cold, Dry Cough, Snoring are taken to consider for predicting the lung cancer. WEKA implements algorithms for data pre-processing, feature reduction, classification such as Naïve Bayes, SVM, RF. The performances of the algorithms for lung cancer disease are analyzed using visualization tools.

Table 4.1 Lung cancer factors

Factors
Age
Gender
Air Pollution
Alcohol use
Dust Allergy
Occupational Hazards
Genetic Risk
Chronic Lung Disease
Balanced Diet
Obesity

Smoking
passive smoker
chest pain
coughing of blood
Fatigue
weight loss
shortness of breath
Wheezing
swallowing difficulty
clubbing of finger nails
Frequent Cold
Dry Cough
Snoring



Filter: None

Current relation
Relation: lung-cancer
Instances: 32
Attributes: 57

Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)

No.	Label	Count
1	1	9
2	2	13
3	3	10

No.	Name
1	class
2	attribute2
3	attribute3
4	attribute4
5	attribute5
6	attribute6
7	attribute7
8	attribute8
9	attribute9
10	attribute10
11	attribute11
12	attribute12
13	attribute13
14	attribute14
15	attribute15
16	attribute16
17	attribute17
18	attribute18
19	attribute19
20	attribute20
21	attribute21

Class: attribute57 (Nom)

Status
OK

Fig: 4.1.1 Upload Dataset

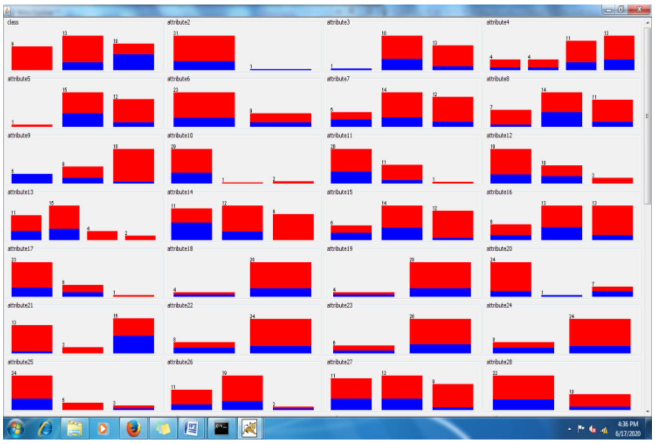
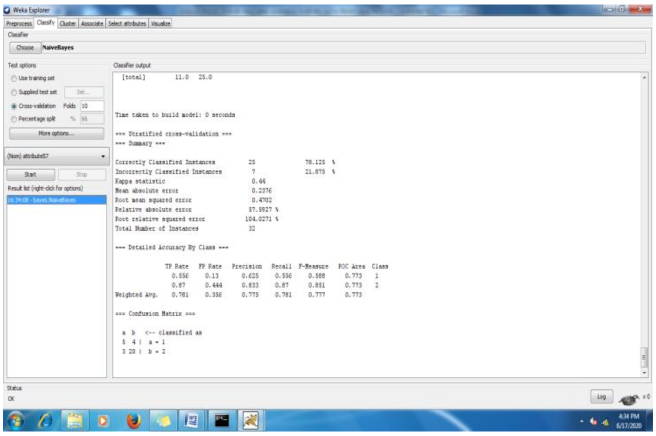


Fig : 4.1.2 Overall Chart



Classifier: NaiveBayes

Test options: Cross-validation Folds 10

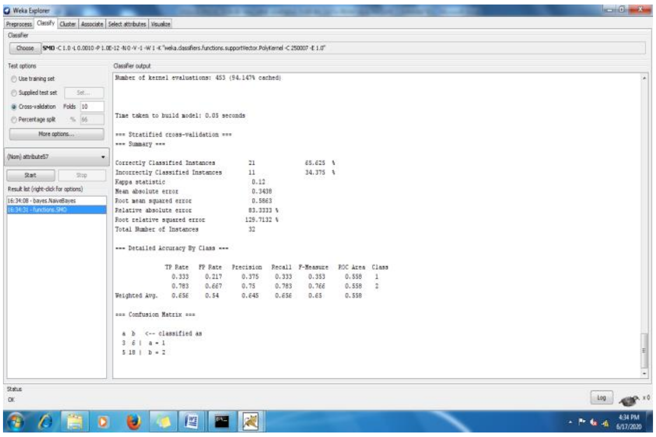
(Nom) attribute57

Result list: 16:34:08 - bayes.NaiveBayes

```
[total]          11.0   25.0



Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          25               78.125  %
Incorrectly Classified Instances         7               21.875  %
Kappa statistic                          0.44
Mean absolute error                      0.2376
Root mean squared error                  0.4702
Relative absolute error                 57.5927 %
Root relative squared error            104.0271 %
Total Number of Instances               32

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.556     0.13       0.625     0.556     0.588      0.773    1
                 0.87      0.444      0.833     0.87      0.851      0.773    2
Weighted Avg.    0.781     0.356      0.775     0.781     0.777      0.773

=== Confusion Matrix ===

  a  b   <-- classified as
  5  4 |  a = 1
  3 20 |  b = 2
```

Status
OK

Fig: 4.1.3 Naive Bayes



Classifier: SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Test options: Cross-validation Folds 10

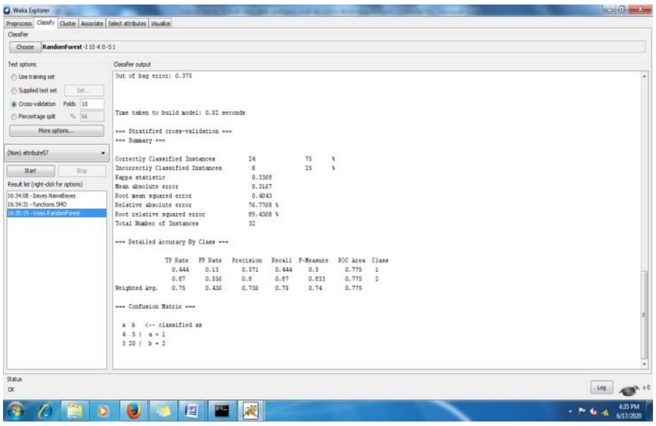
(Nom) attribute57

Result list: 16:34:08 - bayes.NaiveBayes; 16:34:31 - functions.SMO

```
Number of kernel evaluations: 453 (94.147% cached)



Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          21               65.625  %
Incorrectly Classified Instances        11               34.375  %
Kappa statistic                          0.12
Mean absolute error                      0.3438
Root mean squared error                  0.5863
Relative absolute error                 83.3333 %
Root relative squared error            129.7132 %
Total Number of Instances               32

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.333     0.217      0.375     0.333     0.353      0.558    1
                 0.783     0.667      0.75      0.783     0.766      0.558    2
Weighted Avg.    0.656     0.54       0.645     0.656     0.65       0.558

=== Confusion Matrix ===

  a  b   <-- classified as
  3  6 |  a = 1
  5 18 |  b = 2
```

Status
OK

Fig : 4.1.4 Support Vector Machine



Classifier: RandomForest -I 10 -K 0 -S 1

Test options: Cross-validation Folds 10

(Nom) attribute57

Result list: 16:34:08 - bayes.NaiveBayes; 16:34:31 - functions.SMO; 16:35:15 - trees.RandomForest

```
Out of bag error: 0.375



Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          24               75      %
Incorrectly Classified Instances         8               25      %
Kappa statistic                          0.3368
Mean absolute error                      0.3167
Root mean squared error                  0.4043
Relative absolute error                 76.7768 %
Root relative squared error             89.4368 %
Total Number of Instances               32

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.444     0.13       0.571     0.444     0.5        0.775    1
                 0.87      0.556      0.8       0.87      0.833      0.775    2
Weighted Avg.    0.75      0.436      0.736     0.75      0.74       0.775

=== Confusion Matrix ===

  a  b   <-- classified as
  4  5 |  a = 1
  3 20 |  b = 2
```

Status
OK

Fig: 4.1.5 Random Forest

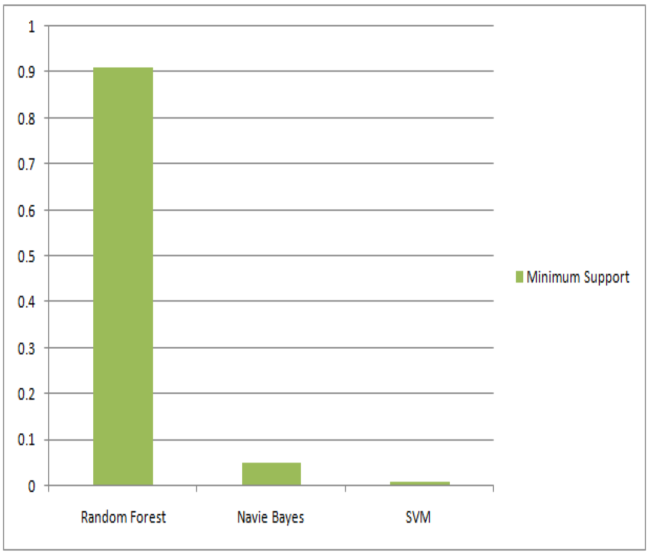
4.2 MINIMUM SUPPPORT

Experiments are performed on the Lung cancer datasets. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were to experiments with WEKA implementations of SVM, Navie Bayes and Random Forest the techniques run to ensure that the results.

TABLE 4.2.1 Showing Confidence for Lung Cancer Dataset

Lung Cancer Dataset	
Technique	**Minimum Support**
Random Forest	0.91
Navie Bayes	0.05
SVM	0.01

CHART 4.2.1 Minimum Support in Dataset



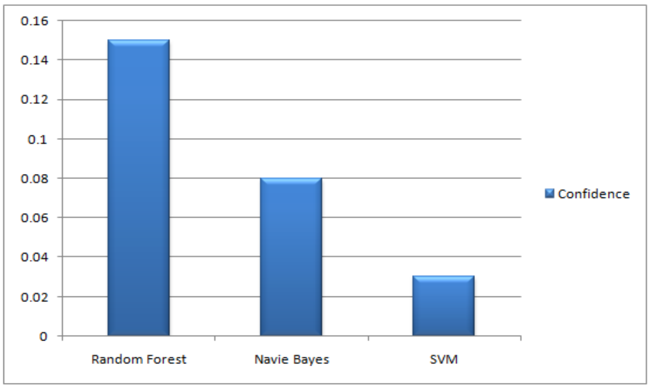
4.2.2 CONFIDENCE

Experiments are performed on the Lung Cancer dataset. Machine with configuration of windows Vista 7 system and 2-GB of RAM is used. The results were to experiments with Weak Tool implementations of SVM,Navie Bayes, Random Forest techniques the run to ensure that the results are comparable for confidence level.

TABLE 4.2.2 Confidence in Lung Cancer Dataset

Lung Cancer Dataset	
Technique	**Confidence Level**
Random Forest	0.15
Navie Bayes	0.08
SVM	0.03

CHART 4.2.2 Confidence in Lung Cancer Dataset



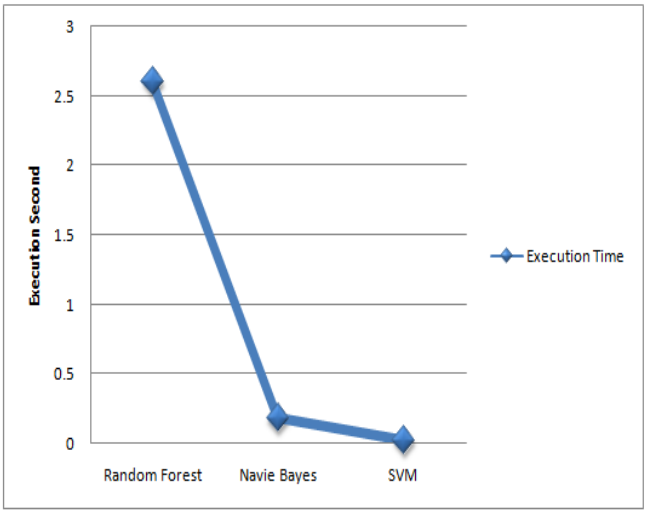
4.2.3 EXECUTION TIME SECOND

Experiments are performed on the datasets Lung Cancer dataset. Machine with configuration of windows Vista 7 system and 2-GB of RAM is used. The results were experiments with Weka Tool implementations of SVM,Navie Bayes, Random Forest the techniques run to ensure that the results are comparable for execution time mille second.

Table 4.2.3 Execution Time Second

Lung Cancer Dataset	
Technique	**Execution Time**
Random Forest	2.60
Navie Bayes	0.18
SVM	0.02

CHART 4.2.3 Execution Time in Lung Cancer Dataset



V. CONCLUSION

In this comparative study SVM, NB,RF techniques are used to find out the support, confidence, time second of Lung Cancer data. High accuracy achieved through Support vector machine technique compare than Naive bayes & RF achieved compare than algorithms. In this paper different algorithm such as Support vector machine (SVM), Naive Bayes, RF. Are experimentally evaluated using lung cancer dataset paper has been carried out. The classification algorithms analyses are based on minimum support, confidence, execution time survey we have known that the best algorithm producing the In future the performance of SVM classifier can be implemented on other datasets also. In future the Naive Bayes algorithm and RF can be hybridized to obtain more effective results.

VI. REFERENCES

[1]. Nikita Jain,Vishal Srivastava, “Data Mining techniques : A survey paper” , International Journal of Research in Engineerning and Technology, pp. 116-119, 2013.

[2]. M.S.B PhridviRaj, C.V. GuruRao, “ Data Mining – Past present and future data streams,” Elsevier, pp. 256-264, 2013.

[3]. K.Kameshwaran, K. Malarvizhi, “Survey on Clustering Techniques in Data Mining,” International Journal of Computer Science and Information Technologies, pp.2272-2276, 2014.

[4]. Gunjan Verma, Vineeta Verma, “Role and Application of Genetic Algorithm in Data Mining,” International Journal of Computer Application, pp. 5-8, 2012.

[5]. Sharaf Ansari,Sailendra Chetlur, Srikanth Prabhu, N. Gopalakrishna Kini, Govardhan Hegde, Yusuf Hyder, “An Overview of Clustering Analysis Techniques used in Data Mining ,” International Journal of Emerging Technology

[6]. Aastha Joshi, Rajneet Kaur, “ A Review: Comparative Study of Various Clustering Techniques in Data Mining,” International Journal of Advanced Research in Computer Science and Software Engineering, pp.55-57,2013.

[7]. Manoj Kumar, Mohammad Husian, Naveen Upreti, Deepti Gupta, Genetic Algorithm “: Review and Application,” International Journal of Information Technology and Knowledge Management, pp.451-454, 2010.

[8]. L.E. Agustın-Blas, S. Salcedo-Sanz, S. Jimenez-Fernandez, L. Carro- Calvo, J. Del Ser, J.A. Portilla-Figueras K. Elissa, “A new grouping genetic algorithm for clustering problems,” Elsevier, pp.9695-9703, 2012.

[9]. Honga Tzung-Pei, Chun-Hao Chenc, Feng-Shih Lin, “Using group genetic algorithm to improve performance of attribute clustering,” Elsevier, pp.1-8, 2015.

[10]. Danial Gomes Ferrari, Leandro Numes de Castro, “Clustering algorithm selection by meta-learning systems: A new distance based problems characterization and ranking

combination methods," Elsevier, pp.181-194, 2015.

[11]. Rajashree Dash and Rasmita Dash, "Comparative analysis of K-means and Genetic algorithm based data clustering," International Journal of Advanced Computer and Mathematical Sciences, pp.257-265, 2012.

[12]. Edvin Aldana-Bobadhilla, Angel Kuri-Morales, "A Clustering based method on the maximum entropy principle," Entropy Article, pp. 151-180, 2015.

[13]. Kannuri Lahari, M. Ramakrishna Murty, and Suresh C. Satapathy, "Study of Classification Algorithm for Lung Cancer Prediction," Advances in Intelligent Systems and Computing," pp. 338, 2015.

[14]. Rahila H. Sheikh, M. M.Raghuwanshi, Anil N. Jaiswal, "Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms," IEEE, pp.314-319, 2008.

[15]. K.Arun Prabha, R.Saranya, "Refinement of K-means clustering using Genetic algorithm," Journal of Computer Application, pp. 256-261, 2011.

[16]. M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", IEEE, pp.580-584, 2014. 17M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", International Journal of Applied Engineering Research, pp. 228-231, 2015.

Cite this article as :

T. Shanmuga Priya, Dr. T. Meyyappan, "Disease Prediction by Machine Learning Over Big Data Lung Cancer", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 16-24, January-February 2021. Available at doi : https://doi.org/10.32628/CSEIT206669
Journal URL : http://ijsrcseit.com/CSEIT206669