International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Prediction of Lung Cancer Using Classifier Models

Shamreen Fathima Saddique[1], Sharmithra P[2], Justin Xavier D[3]

Student, Dept. of Computer Science, Mohamed Sathak A.J College of Engineering, India[1]
Student, Dept. of Computer Science, Mohamed Sathak A.J College of Engineering, India[2]
Asst. Professor, Dept. of Computer Science, Mohamed Sathak A.J College of Engineering, India[3]

Abstract: ***In recent years, Lung Cancer has become a serious disease that threatens the health and mind of human. Efficient predictive modeling is required for medical researchers and practitioners. This study proposes a lung cancer prediction model based on naïve Bayes which aims at analyzing some readily available indicators (age, smoking, alcohol consumption, chest pain, etc.) effects on lung cancer and discovering some rules on given data. The method can significantly reduce the risk of disease through digging out a clear and understandable model for lung cancer from a medical database. Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The validation of results at Chennai Port Hospital shows that the naïve Bayes algorithm can greatly reduce the problem and it can effectively predict the impact of these readily available indicators on the risk of lung cancer. Additionally, we get a better prediction accuracy using naïve Bayes than the support vector machine algorithm, logistic regression and random forest.***

Keywords: prediction model; naïve Bayes; Lung Cancer

1. Introduction

Machine learning is a field of computer science that gives computer systems the ability to "learn" (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed. machine learning explores the study and construction of algorithms that can learn from and make predictions on data– such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible.

RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Windows, macOS, and Linux.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, macOS, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro. Further Rstudio allows us to perform predictive analysis using machine learning algorithms.

2. Problem Definition

Modern medicine generates a great deal of information stored in medical database. In today's world, every individual is facing growing health issues which need to be cured quickly. With continually increasing lung cancer in patients due to high intake of tobacco and puff, predicting the cancer in patients at an early stage is the huge issue for the clinicians to make decisions. Since it is considered as a taboo in some countries people fear to come forward to diagnose the disease, the best place to find occurrence of disease is by applying machine learning concept to create the predictive model by using the data collected in the hospital regarding the patients affected by lung cancer to predict lung cancer.

3. Existing System

In the existing system, the model is created for the prediction of Lung Cancer. It is based on the basis of Support Vector Machine Algorithm. The model is established on the CT images.

Drawbacks

- The existing model requires scans images of the patients to detect the presence of cancer in patients.
- 77% of accuracy is achieved in this model.
- The model is trained with only 12 CT scans in which 6 of them are cancer affected patients and rest are non-affected.

4. Proposed System

4.1 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, Yellow Finger, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Alcohol Consumption, Smoking, chest pain, coughing of blood, shortness of breath, wheezing, swallowing difficulty are taken to consider for predicting the lung cancer. Rstudio implements algorithms for data pre-processing, feature reduction, classification such as Naive Bayes, Random Forest, Support Vector Machines, K-Nearest-Neighbours are also implemented. The performances of these algorithms for lung cancer disease are analyzed using confusion matrix.

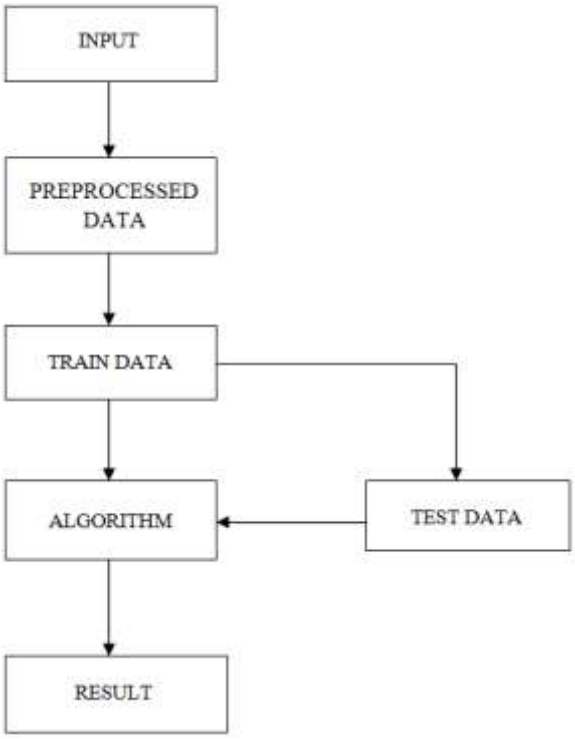
Table 4.1: Lung Cancer Factors

Attributes
AGE
GENDER
YELLOW FINGER
ANXIETY
PEER PRESSURE
SMOKING
CHRONIC LUNG DISEASE
ALCOHOL CONSUMPTION
WHEEZING
COUGHING OF BLOOD
FATIGUE
SWALLOWING DIFFICULTY
SHORTNESS OF BREATH
CHEST PAIN
ALLERGY

4.2 Model Description

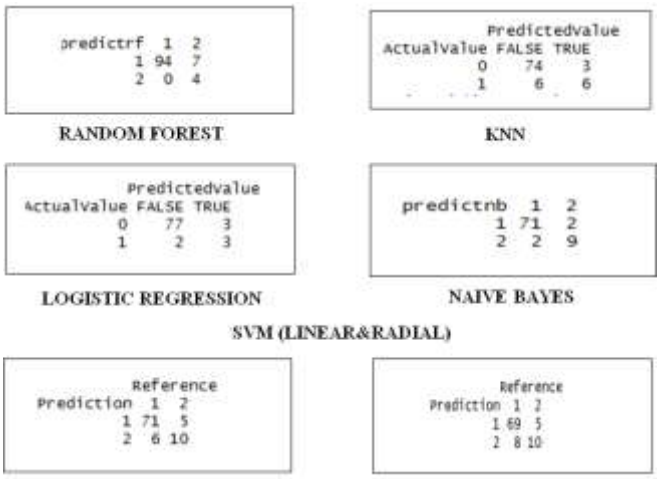
The given raw datasets are processed to remove any outliers to form a preprocessed data in order to achieve an higher accuracy in a prediction of lung cancer models. Meanwhile the data is also checked for any missing values and is also normalized to reduce the error rate. The preprocessed data is further divided into two datasets with 70% of its data belonging to train datasets and the rest 30% is the test dataset.

In the proposed system, various prediction models are created using the train dataset based on algorithm such as naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, KNN. Further the test dataset are given to these models and the confusion matrix is created for each of the following models such as naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, KNN model to evaluate the performance accuracy of these models.



Flow Chart 4.2.1: Proposed Model

5. Performance Analysis



Screenshots 5.1: Confusion Matrix

The above screen shots describes the performance analysis of the algorithm and compares the accuracy score of the algorithms such as naïve Bayes, KNN, Support Vector Machine, Random Forest . The normal accuracy of the existing system is 77% and for proposed system, the higher accuracy achieved is 95.24% using naïve Bayes and almost 88% for the rest of the algorithms. The performance is analyzed and gives the result for higher accuracy in prediction of Lung Cancer.

Table 5.2: Performance Accuracy

NAÏVE BAYES	95.24%
RANDOM FOREST	93%
KNN	89.90%
LOGISTIC REGRESSION	88.53%
SVM-LINEAR	88.04%
SVM-RADIAL	85.87%

6. Conclusion and Future Work

This paper introduces the Machine Learning in health care management is not analogous to the other fields due to the reason that the data existing here are heterogeneous in nature and that a set of ethical, legal, and social limitations apply to private medical information. The experiment has been performed using Rstudio tool with several machine learning classification algorithm and it is found that the Naive Bayes algorithm gives a better performance over the other classification algorithm such as Support Vector Machine, Random Forest, Logistic Regression, KNN. Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other machine learning algorithms.

References

[1] Weifeng Xu, Jianxin Zhang, Qiang Zhang*, Xiaopeng Wei (2017),” Risk prediction of type II diabetes based on random forest model”, IEEE-2017.
[2] Dr. S. Vijayarani, S. Dhayanand (2015), “Liver Disease Prediction using SVM and Naïve Bayes Algorithms”, IJSETR-2015.
[3] Dania Abed Aljawad, Ebtesam Alqahtani, Ghaidaa AL-Kuhaili, Nada Qamhan, Noof Alghamdi, Saleh Alrashed,

Jamal Alhiyafi, Sunday O. Olatunji8" Breast Cancer Surgery Survivability Prediction Using Bayesian Network and Support Vector Machines" IEEE-2017

[4] N. Ramkumar, Dr. S. Prakash, S. Ashok kumar, K Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem" IEEE-2017