JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTION OF LUNG CANCER USING SUPERVISED MACHINE LEARNING ALGORITHMS

[1]Swarnalatha Prathipati,[2]Yada Roshik,[3]Nandini Shailesh Kadre,[4]Kolla Abhishek and [5]Kommera Nikhil Reddy
[1]Assistant Professor,[2,3,4,5]Student
[1] Dept. of CSE, GITAM University, Visakhapatnam, India

ABSTRACT: Lung cancer is the leading cause of cancer-related death in this generation and is expected to remain so for the foreseeable future. If lung cancer signs are identified early, the disease may be treatable. The main factor which causes lung cancer is smoking. Various machine learning algorithms can be used to detect lung cancer. In this study, we are using various machine learning algorithms like logistic regression, support vector machine (SVM), k-Nearest Neighbor (KNN), Random forest, decision tree algorithm to detect what type of cancer low, medium, high i.e., low depicts no cancer, medium depicts benign, high depicts malignant is present across different age groups like Youth, Working Class and Elderly. This study will also focus on how smoking affects various. The classification models will be generated using the training data and the corresponding models will be evaluated using the test data to obtain the accuracy of the models. Finally, we will compare the accuracy rates of each classification model we will implement based on metrics such as F1 score, recall, precision, specificity and arrive at a conclusion.

Keywords--- ***Machine Learning; Decision Trees; Lung Cancer; Non-Small Cell Lung Cancer (NSCLC);Feature selection; Small Cell Lung Cancer (SCLC); Feature Importance; Data Preprocessing; Bootstrapping***

1.INTRODUCTION

Lung Cancer plays a vital role. The lung cancer is a type of cancer which begins in lungs.
Lungs are most useful organs in human body where these are used to inhale of oxygen and exhale of carbon-dioxide and these are spongy like structured in our chest. There many types in cancers like Breast Cancer, Prostate Cancer, Lung Cancer, Bladder Cancer, Kidney Cancer, Liver Cancer, Pancreatic Cancer, Thyroid Cancer. In the majority of cancer's people are suffering from lung cancer.

The two histological subgroups of lung malignancies are non-small cell lung cancer (NSCLC) (80.4%) and small cell lung cancer (SCLC) (16.8%).[4] Despite the fact that the precise mechanism behind lung cancer have not yet been fully understood, a number of factors, such as cigar smoke, ionising radiation, and viral infection, it may play a major role in the emergence of lung cancer.

Lung cancer, the second-most hazardous disease in the world, is recoverable if detected early.
In furthermore, ongoing research is being done in this realm of lung cancer detection to achieve 100% detection accuracy.[5] The characteristics of the tumour termed as prognostic factors for cancer affect and forecast the likelihood that cancer patients will survive.[6] Lung cancer can occur in two different ways. They are Non-Small cell lung cancer (NSCLC) and the Small cell lung cancer (SCLC). The SCLC is also called as the oat cell.

According to a research from 2021, there were 1.9 million cancer patients in India, with the most prevalent six types being breast, lung, pancreas, ovary, colon-rectum, and stomach. The graph in [10] depicts about the analysis of the most typical cancer patients. The majority of cancer patients were examined at a nearby radical phase for the above six cancers mentioned.

2.RELATED WORK

In a paper where we analyzed symptoms using machine learning for early stage lung cancer authors Atharva Bankar, Kewal Padamwar, Aditi Jahagirdar rather than analyzing symptoms and their lifestyles using text-based data, this study helps doctors and other health professionals more effectively treat and diagnose lung cancer patients. In my research, some algorithms such as Decision Trees, Random Forest, XG Boost achieved 100% accuracy on the data set they used, and these 3 algorithms gave 100% accuracy in the youth and working age groups. active and 93% in the elderly group. . lung cancer prediction groups.[1]

In this paper they have used 2 Algorithms SVM, Random Forest the cancer is a very important disease that claims many lives worldwide, we focused on cancer prediction in this study. The three forms of cancer mentioned in [2] are lung cancer, breast cancer, and prostate cancer. Breast cancer prognosis is significant in the Medicare and biomedical fields.[2]

The main purpose of this research conducted by the authors was to analyze and compare the results of multi-layer perceptrons, neural networks, decision trees, naive Bayes, gradient-boosted trees, support vector machines, random forests, and majority voting. Gradient-boosted trees were found to outperform all other classifiers used in studies using K-fold cross-validation on the University of California, Irvine lung cancer dataset. [7]

The main goal of [8] is that cancer is a condition in which the body's cells grow out of control. Lung cancer is the name given to cancers that first appear in the lungs. Lung cancer can also start in the lungs, in addition to other organs, including lymph nodes and the brain. Lung cancer can spread from other organs. Metastases are the term used to describe the spread of cancer cells through one organ to another. In this study, lung cancer is predicted using GNB machine learning techniques. Using the University of California, Irvine Machine Learning Repository, the suggested GNB algorithm's performance is assessed. Performance investigation reveals that GNB's prediction model outperforms other machine learning methods by 98%.

3.METHODLOGY

The methodology is described in Fig.1 at first we have identified the problem and then we have gathered the data i.e., dataset from the [3] and after the data gathering we have done data preprocessing It is the process of preparing data for machine learning is critical, involving the conversion of unrefined information into a structure that can be utilized to coach models. The objective behind this process is to format data in such a way as to bestow effective knowledge upon our model.

There are some Preprocessing techniques:-

1.Data Cleaning:- To complete this task, it is necessary to eradicate any absent or inaccurate information from the dataset.

2.Data Normalization:- The process consists of adjusting the data so that it falls within a uniform scope or arrangement. This is done to guarantee that each aspect shares comparable scopes and strengthens certain machine learning algorithms' functions.

3.Data Encoding:- The process entails the transformation of types of information into numeric values that are suitable for utilization in machine learning frameworks. Such a conversion may be accomplished through diverse approaches such as employing methodologies like one-hot encoding or label encoding.

4.Feature Selection:- To increase model accuracy, it is crucial to select the most pertinent traits from all available data. This process involves reducing dimensionality and optimizing feature selection techniques.

5.Data Splitting:- The process entails the partitioning of collected information into distinct training and testing sets, which is indispensable in assessing how well a model can perform when exposed to unprecedented data.

Data visualization it is the next step after the data preprocessing it is Visualizing data is the art of representing information through graphics. This entails transforming sets of numerical or qualitative data into visually appealing forms such as diagrams, scatterplots and heat maps to facilitate comprehension. Utilized in analyzing information for knowledge pattern recognition, trend identification and correlation discovery between variables; visual representation is a necessary tool used by professionals across various industries seeking insight from vast amounts of collected data.

Feature Extraction in this step The act of feature extraction involves the selection and transformation of vital data into a set of features that are pertinent to analysis or modeling. This process revolves around reducing dataset dimensionality by choosing crucial elements and transforming them into a more significant representation, which is compact yet meaningful. To put it differently, this activity entails taking the most relevant information from raw data for further processing purposes.

Model Building is the art of machine learning involves crafting a numerical blueprint for analyzing the information, known as model building. This scientific method entails employing algorithms and datasets to create an accurate portrayal of a system or problem. The end goal is to produce a prognostic pattern that empowers individuals to make knowledgeable choices derived from new data results. The typical progression for creating a model requires several stages they are Data preparation, selection of features, selection of algorithms, training of models, evaluation and tuning of models, and deployment of models.

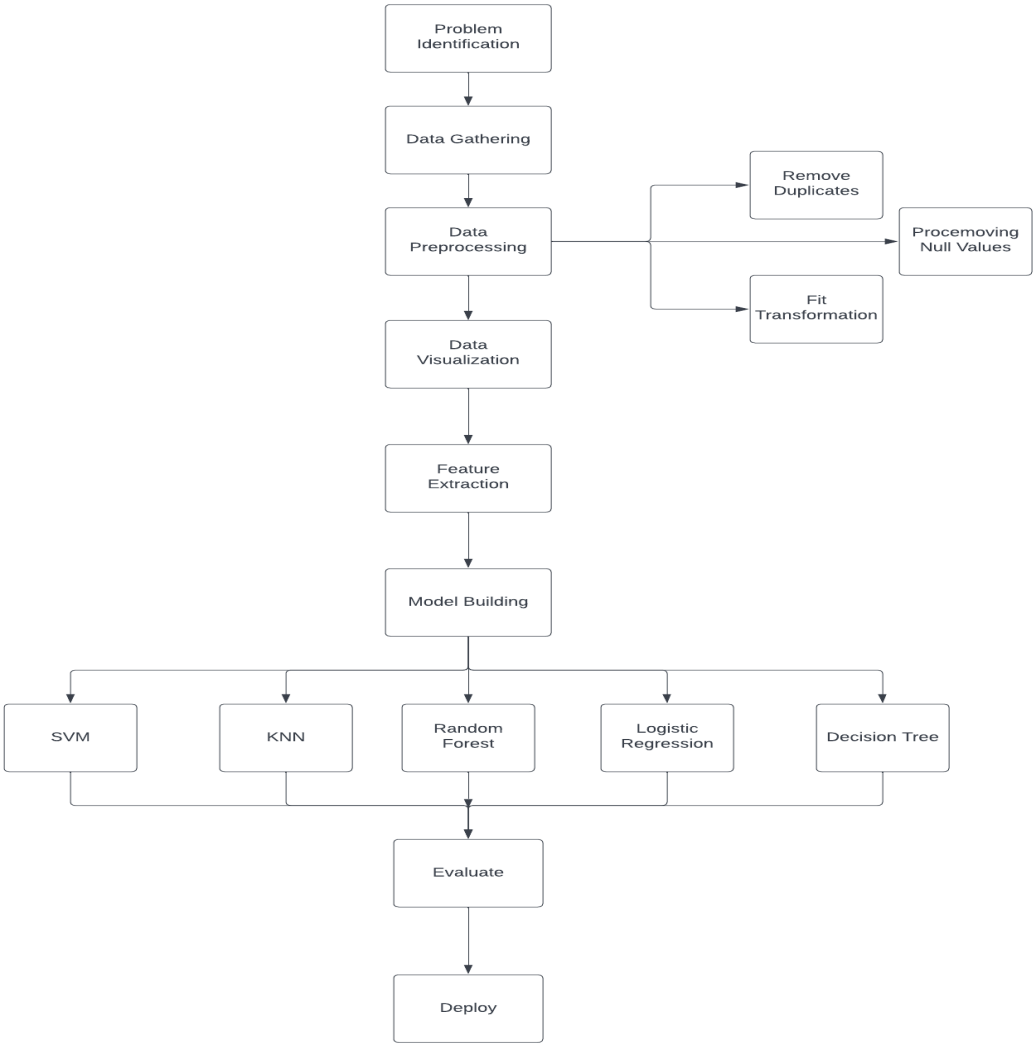

Fig. 1. workflow

4.DATASET / PROPOSED WORK

The data used in this paper to train the models were obtained using the dataset "Lung Cancer Data" from the data.world's global data catalogue. There are 1000 instances and 25 properties per instance. (With 24 independent variables and 1 dependent variable). The data's primary component ordinal values make it ideal for performing a relative analysis on the independent variables.[1]

The dataset's attributes are listed as follows:-
Patient-ID, Chronic Lung Diseases, Passive Smoker, Weight Loss, Swallowing Difficulty, Snoring, Age, Dust Allergy, Balanced Diet, Chest Pain, Shortness of Breath, Clubbing of Finger Nails, Alcohol Use, Gender, Occupational Hazards, Obesity, Coughing of Blood, Wheezing, Frequent Cold, Air Pollution Index, Genetic Risks, Smoking, Fatigue, Dry Cough, Level [1,3]

A.SVM

Machine learning algorithms such as SVM, also known as Support Vector Machines, have become increasingly popular for tasks such as classification and regression. SVMs are based on the idea of finding the best hyperplane that separates data into different classes. A good generalization is achieved by choosing a hyperplane that maximizes the margin between classes. A SVM algorithm divides training data points into classes based on the best hyperplane that separates the vectors in the high-dimensional space. In order to achieve good generalization performance, the hyperplane is chosen in such a way as to maximize the margin between the classes.

B.KNN

Amongst the plethora of machine learning techniques, K-Nearest Neighbor stands out as a fundamental algorithm that relies on supervised learning. Its implementation involves intricate calculations and complex decision-making processes while operating within vast datasets. Assuming that the new example and the prior cases are comparable, the K-NN method places the new instance in the category that most closely resembles the present categories. The KNN approach simply stores the data during the training phase and summarises real time data into a category that is very similar to the training data.

C.RANDOM FOREST

The Random Forest algorithm is versatile, handling both classification and regression tasks. It employs ensemble learning by creating many decision trees to create a more accurate model. In this method, numerous randomized subsets of training data and features are

utilized in each generated decision tree; such randomness reduces overfitting while elevating the generalization ability of the final model. To build these trees during training requires bootstrapping which consists of randomly taking multiple samples from provided data with replacements made along the way. Given that every trained tree uses different selected sections within one piece set aside for sampling purposes means predictions will be mixed-and-matched resulting in an overall prediction combination pulled from all built models created beforehand without interaction through varying stages when running or making new project resolutions because everything works together collaboratively as intended throughout implementation stages alike!

D.LOGISTIC REGRESSION

The algorithm of Logistic Regression is utilized in machine learning for the purpose of tackling binary classification tasks. These types of tasks are concerned with predicting a specific outcome that falls under two categories, such as yes or no and true or false. The model stands out among the various linear models due to its utilization of a logistic function which allows it to predict probability pertaining to outcomes better than other methods used by different linear models.Logistic regression is a method in which the features are combined linearly to create a logit. This created value then goes through transformation using logistic function, resulting in an outcome of probability ranging from 0 and up to 1. The input can be considered as real-valued since it undergoes mapping with logistic function converting values between zero or one that indicates class positivity likelihoods.

E.DECISION TREE

A machine learning algorithm called Decision Trees is quite popular in performing classification and regression tasks. This type of supervised learning approach learns a set of hierarchical decisions based on the input features to predict what variable it's targeting. The method partitions your data recursively into smaller subsets founded solely upon its value within an input feature at each given node throughout this tree-like model structure: the best separator for these classes or groups is then chosen through metrics such as metric impurity or gain ratio, continuing until pure sets exist among leaves that target variables can be found inside them almost exclusively without other factors being present in either class's composition - regardless if they are less than completely perfect matches!

F.FEATURE IMPORTANCE

The effort of this study is to gain understanding on the analysis of triggers that have a direct influence within an array of age groups, acknowledging variations in their lifestyles. One notable finding was the linkage between advancing years and prevalence rates for lung cancer as discussed by [9]. The materials are sorted into three sections with reference to particular age brackets, aimed at carrying out a comparative assessment pertaining symptoms. Table-1 herein displays such groupings distinctively.

Table-1

The Age Groups and Number of Instances

Age Groups	Youth: 0-25	Middle Age: 25-50	Old age: 50-80
No. of Instances	165	700	134

Within the realm of machine learning, performance measurement for models is accomplished through evaluation metrics. Tasks involving classification rely heavily on several commonly used evaluation metrics such as:

1.**Precision** is the measure of how frequently a model's predictions are correct compared to all its estimations. The accuracy of a prediction is measured as a percentage of all model-induced forecasts.

2.**Accuracy** can be determined by calculating the proportion of true positive predictions to all positive predictions made. This calculation measures how accurately a model has predicted positivity in relation to its overall number of prognostications.

3.Remembering the concept, one calculates **Recall** by dividing the true positive predictions with all actual instances that are positively identified in a dataset. This metric is indicative of how effective a model performs when it tries to detect such specimens.

4.The score known as **F1-score** it is calculated by taking the reciprocal of the sum of one divided by precision and one divided by recall, then multiplying that result with two. This provides a harmonious balance between these two metrics to get an accurate measure of performance in a single value. It's commonly used because it takes into account both false positives (precision) and false negatives (recall).

5.**Support** Assistance is determined by the total quantity of occurrences within each category present in a dataset.

One can implement a variety of measuring criteria to evaluate the efficiency of classification models and contrast varying installations. To obtain an extensive comprehension of how adequate a model operates, it is crucial to examine numerous assessment metrics concurrently.

5.RESULT AND ANALYSIS

Table-2 demonstrates about the accuracy of models based on trees performs better across the entire dataset. We are able to locate and choose the most crucial features that help patients be identified as having lung cancer thanks to these tree-based models' regulations operate the metrics.

Table-2
The Performance Metrics of:-
Accuracy, Recall, Precision, Support, F1-score

Algorithm	Accuracy	Recall	Precision	Support	F1-Score
SVM	78%	0.77	0.77	250	0.77
KNN	99.2%	0.99	0.99	250	0.99
Random Forest	100%	1	1	250	1
Logistic Regression	87.6%	0.87	0.87	250	0.87
Decision Tree	100%	1	1	250	1

6.CONCLUSION AND FUTURE SCOPE

With the help of this research, we come to know how the lung cancer is varying across different age groups and what causes lung cancer across various age groups. With the accuracy of data and using various machine learning algorithms we can conclude that if the data is reliable we can get good accuracy. The accuracy of SVM algorithm is 78% which is the least whereas the Random forest and Decision Tree algorithms have an accuracy of 100% which is the highest accuracy. These algorithms will be useful for the doctors to analyse the symptoms and causes of lung cancer.

The future scope of the study includes predicting cancer through CT scan images and use of deep learning algorithms for more enhanced and detailed analysis of lung cancer.

REFERENCES

1. “Symptom Analysis using a Machine Learning approach for Early Stage Lung Cancer” by Atharva Bankar, Kewal Padamwar and Aditi Jahagirdar.
2. “Cancer Prediction using Machine Learning” by Ganta Sruthi, Chokkakula Likitha Ram,Malegam Koushik Sai, Bhanu Pratap Singh, Nikhil Majhotra, Neha Sharma
3. Data set from https://data.world/cancerdatahp/lung-cancer-data website.
4. Travis WD, Travis LB, Devesa SS. Lung cancer. *Cancer*. 1995;**75**(1 Suppl):191–202.
5. A Study On Prediction Of Lung Cancer Using Machine Learning Algorithms
Abhishek Gupta, Zuha Zuha ,Israr Ahmad, Zeeshan Ansari https://doi.org/10.21203/rs.3.rs-1912967/v1
6. C-Q Zhu, W Shih, C-H Ling, M-S Tsao 2006.Immunohistochemical markers of prognosis in non-small cell lung cancer: a review and proposal for a multiphase approach to marker evaluation. Journal of Clinical Pathology2006;59:790-800
7. M. I. Faisal, S. Bashir, Z. S. Khan and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), Karachi, Pakistan, 2018, pp. 1-4, doi: 10.1109/ICEEST.2018.8643311
8. Anita, C. S., Vasukidevi, G., Rajalakshmi, D., Selvi, K., & Ramesh, T. (2022). Lung cancerprediction model using machine learning techniques. International Journal of HealthSciences, 6(S2), 12533–12539. https://doi.org/10.53730/ijhs.v6nS2.8306
9. R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001
10. ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS USED TO DETECT LUNG CANCER ANJALI RAJ, AMBILY JACOB https://ijcrt.org/papers/IJCRT22A6139.pdf or IJCRT22A6139.pdf