International Journal of Electrical and Electronics Engineers
Volume 14, Issue No. 01, Jan-June 2022
ISSN (O) 2321-2055
ISSN (P) 2321 -2045

LUNG CANCER PREDICTION USINGMACHINE LEARNING

Batch-13: P. Mounika, N. Yamuna, SK. Ameen Rehamath, S. Dileep KumarUnder the Guidance of Mrs. B. Thriveni M. Tech,

Assistant Professor,

Department of Electronics and Communication EngineeringTirumala Engineering College, Narasaraopet.

Abstract:

Lung cancer disease cause of cancer death in both men and women, and it is the second most common cancer in the world, after skin cancer. Lung cancer claims more lives each year than breast, collector, ovarian, and prostate cancer combined. Its mortality rate is three times that of prostate cancer deaths and nearly double that of breast cancer deaths in women. Lung cancer is now responsible for 32% of cancer deaths in males and 20% of cancer deaths in women. Lung cancer symptoms can take years to appear, and in many cases, there are no signs at all until the disease is advanced. Lung cancer symptoms are frequently misdiagnosed as less serious issues. Alternatively, they are assumed to be solely tied to tobacco usage in smokers. Each year, around 222,000 new instances of lung cancer are identified in the United States. We present Stratified K-Fold Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbour's (KNN), Gaussian Naive Bayes (GNB) classifiers for lung cancer prediction based on symptoms in this research. Finally optimize the accuracy of the result data. The performance of the techniques is evaluated based on accuracy and precision. Then processing of some of the attributes provided identifies the existence of lung cancer and provides the graphical model visualization.

Keywords: Support Vector Machine, Random Forest, Stratified K-Fold, Lung cancer

1.Introduction:

Cancer is a term refers to a large group of diseases that can damage any body parts. malignant tumors and neoplasms are other terms that can be used instead. A characteristic of cancer is that abnormal cells can form rapidly, grow beyond normal limits, and invade adjacent parts of the body and spread to other organs. The latter process is called metastasis. Widespread metastasis is the leading cause of death from cancer. Lung cancer continues to be the world's deadliest and most expensive disease.

It has a three-fold risk of mortality compared to prostate cancer deaths. Mortality from prostate cancer is roughly twice as high as breast cancer deaths. Lung cancer is the primary cause of cancer death in the United States. Men account for 20% of cancer mortality, whereas women account for 20%. Lung cancer is affected by many factors, such as environmental and behavioral factors, etc. Smoking is the most important thing. Smokers are 20 times more likely to develop lung cancer than non-smokers. Professional Exposure to carcinogens (asbestos, arsenic, and other carcinogens), aromatic hydrocarbons, fuel pollution, underlying lung disease, and a family history of lung disease Cancer, nutritional factors (vitamins A, B, and C), substance abuse, and use are the other main risk factors for lung cancer. These factors vary by race, country, and region. Other Factors Affecting Lung Incidence and Mortality Cancer is the Human Development Index (HDI). Alternatively, the index is based on quality of life, health and health facilities, fearlessness, relaxation, the economy, and social security.

2. Literature Survey:

Lung Cancer prediction using Machine Learning Algorithms to find out the accuracy. The properties for a good accuracy are:

1. Should give the high accuracy.
2. Should predict lung cancer quickly.
3. Should classify the data properly.

Outlier prediction is a critical task as outliers indicate abnormal running conditions from which significant performance degradation may happen. Techniques used in prediction can be divided into two:

a. Supervised techniques where past known data are used to build models which will Produce a suspicion result for the new data.

b. Un supervised are those where there are no prior sets in which the set of the data is known to new data.

3. Existing System:

This was on k nearest neighbors' algorithm implementation, only the two features with the most variance was used to train the model.

The model was set to have predict the lung cancer is a highly researched field, there are many different algorithms and techniques for performing the prediction system.

Some other various existing algorithms used in the lung cancer prediction includes Gaussian Naive Bayes, support vector machine (SVM), K-Nearest Neighbours, Naïve Bayes etc.

3.1 k-Nearest Neighbor:

K- Nearest Neighbors (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms.

KNN is used in a variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers.

Loan disbursement, banking institutes will predict whether the loan is safe or risky. In Political science, classifying potential voters in two classes will vote or won't vote.

KNN algorithm used for both classification and regression problems.

KNN algorithm based on feature similarityapproach.

3.2 Super Vector Machine:

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression, and outlier detection.

The advantages of support vector machines are:

1. Effective in high-dimensional spaces. Still effective in cases where the number of dimensions is greater than the number of samples.

2. Uses a subset of training points in the decision function (called support vectors).

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

3.3 Naive Bayes:

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Suppose we have a dataset of weather conditions and corresponding target variable "Play". So, using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions.

So, to solve this problem, we need to follow the below steps:

Convert the given dataset into frequency tables.

1. Generate Likelihood table by finding the probabilities of given features.
2. Now, use Bayes theorem to calculate the posterior probability.

3.4 Random Forest:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the data set and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or the same algorithm multiple times to form a more powerful prediction model.

The random forest algorithm combines multiple algorithms of the same type i.e., multiple decision trees resulting in a forest of trees, hence the name "Random Forest".

The random forest algorithm can be used for both regression and classification tasks.

4. Proposed System:

Our goal is to implement a machine learning model in order to classify, to the highest Possible Degree Of accuracy, lung cancer prediction from a data set gathered from Kaggle.

After initial data exploration, we knew we would implement a logistic regression model for best accuracy reports. Gaussian naive Bayes, Random Forest as they are good candidates for binary classifications, python skit learn library was used to implement the project, we used Kaggle datasets for lung cancer prediction, using pandas' data frame for class==0 for no cancer and class==1for plotting the cancer and non-cancerous data

4.1 K-Fold Cross Validation Technique:

In each set (fold) training and the test would be performed precisely once during this entire process. It helps us to avoid overfitting. As we know when a model is trained using all of the data in a single short and give the best performance accuracy. To resist this k-fold cross-validation helps us to build the model is a generalized one.

To achieve this K-Fold Cross Validation, we have to split the data set into three sets, Training, Testing, and Validation, with the challenge of the volume of the data. Here Test and Train data set will support building model and hyperparameter assessments.

In which the model has been validated multiple times based on the value assigned as a parameter and which is called K and it should be an INTEGER.

Make it simple, based on the K value, the data set would be divided, and train/testing will be conducted in a sequence way equal to K time.

5.Result:

Predict

You are not showing any signs of lung cancer

Predict

You may affect with lung cancer. Please contact a doctor

6. Conclusion:

The Random Forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre- processing techniques would also help.

The SVM algorithm still suffers from the imbalanced data set problem and requires more pre-processing SVM is great but it could have been better if more to give better results at the results shown by SVM is great but I could have been better if more pre -processing have been done on the data.

7. Future Enhancement:

We design a system to predict the lung cancer This system is capable of providing most of the essential features we require to predict lung cancer.

We have just analysed the probability of having lung cancer not suggesting any treatment. Predicting lung cancer without any tests in real time is not easy but it is feasible by doctors. The proposed architecture is basically designed to predict the lung cancer.

Future enhancement can be done by making the prediction of lung cancer more accurate by improving algorithms and collecting. By collecting a greater number of data sets. We can also suggest a lab to collect samples of that person and go through testing. After that testing the reports were scanned and analysed using nodal analysis and then person is having cancer or not. confirms that the person is having cancer or not

through testing. After that testing the reports were scanned and analysed using nodal analysis and then confirms that the person is having cancer or not

8. References:

J. Amin, M. Sharif, M. Raza, T. Saba and M. A. Anjum, “Brain tumour detection using statistical and machine learning method,” Computer Methods and Programs in Biomedicine, vol. 177, pp. 69–79, 2019.

A. M. Salem, “Advances in intelligent analysis of medical data and decision support systems,” in Machine Learning Applications in Cancer Informatics. Berlin, Germany: Springer, pp. 1–14, 2013.

Y. Gültepe and N. Gültepe, “Preliminary study for the evaluation of the haematological blood parameters of seabream with machine learning classification methods,” The Israeli Journal of Aquaculture Bamidgeh, vol. 72, pp. 1–10, 2020.

S. Marshland, Machine Learning: An Algorithmic Perspective, 2nd ed., Boca Raton, Florida, USA: Chapman & Hall/CRC, 2009.

P. G. Espejo, S. Ventura and F. Herrera, “A survey on the application of genetic programming to classification,” IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, vol. 40, no. 2, pp. 121–144.