LUNG CANCER PREDICTION USING PREDICTIVE ANALYSIS

**DESCRIPTION:**

This model is designed for predicting lung cancer using textual data. Textual data in the sense ,we investigated about sex, age , smoking habit ,alcohol consumption, allergy , wheezing, chronic disease , fatigue , yellow finger , chest pain etc. This model will predict the lung cancer based on accuracy.

In existing system **,** the lung cancer prediction which are being done through the Chest X-Ray, Magnetic Resonance Imaging (MRI) scan and computed tomography(CT) scans, PET(positron emission Tomography).This scan can sometimes detect disease before it shows up on other imaging tests and Bronchoscopy etc., by the health profession.

Planning to build the model in Python.

**TECHNOLOGY DETAILS:**

RapidMiner Studio

Algorithms: Naïve Bayes and Logistic Regression

**SCREEN DETAILS**:

1. Data collection module
2. Data preparation module
3. Apply model
4. Training module
5. Model Evaluation and testing module
6. Predictive analysis
7. **Data collection module:**

Data collection is used for gathering and measuring information from variety of sources to get a complete and accurate dataset .

Our dataset contains numerical values 1 and 2.

1= patient is affected with the symptom

2= patient is not affected with the system

T= true

F = False

Our dataset contains the following fields:

|  |  |  |  |
| --- | --- | --- | --- |
| Patient ID | Alcohol habit | Chronic disease | Chest pain |
| Age | Peer pressure | coughing | Yellow fingers |
| Gender | fatigue | Breathing trouble | wheezing |
| Smoking habit | allergy | Swallowing problem | Lung cancer (T/F) |

1. **Data preparation module:**

This module is used for cleaning and transforming raw data prior to processing and analysis. It is to ensure that data is consistent and of high quality.

This is used for removing the unwanted values and missing labels.

1. **Apply model;**

The data collected and fed into the tool to find accuracy of algorithms. We use supervised machine learning algorithm such as Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree etc. We will then find out the accuracy of all the supervised learning algorithms and combine two algorithms which have better accuracy.

1. **Training Module:**

The dataset will splitted into training part and testing part. Training part is used to train the model basically to fit the parameters. Test data is used only to assess performance of the model.

1. **Model Evaluation and testing module:**

Model evaluation helps to find out the best model that represents our data and how well the chosen model will work. We use Naïve Bayes (63.3%) and Logistic Regression (51.25%) because the accuracy is better when comparing with other supervised learning algorithms.

1. **Predictive analysis:**

Here we propose a new algorithm as combined Naïve Bayes and Logistic Regression that will be used for predicting lung cancer with real time data collected .

**CURRENT ANALYSIS REPORT:**

**ACCURACY:**

Naïve Bayes = 63.3%

Logistic Regression = 51.25%