PROJECT REPORT

(Project Term January-May 2022)

Lung Cancer Predictions

Submitted by

Abhishek Kumar Singh **11907788**

INT 247

(B. Tech CSE)

Under the Guidance of

Dr. Sagar Pande

School of Computer Science and Engineering

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

DECLARATION

We hereby declare that the project work entitled Lung Cancer Predictions is an authentic record of our own work carried out as requirements of Project for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of Sagar Pande, during January to May 2022. All the information furnished in this project report is based on our own intensive work and is genuine.

Name of Student : Abhishek Kumar Singh
Registration Number: 11907788

CERTIFICATE

This is to certify that the declaration statement made by is correct to the best of my knowledge and belief. They have completed this Project under my guidance and supervision. The present work is the result of their original investigation, effort, and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara.

Name of the Mentor: Dr. Sagar Pande

School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab.

ACKNOWLEDGEMENT

We take this opportunity to express our deep gratitude and most sincere thanks to our teachers, parents, and friends for giving most valuable suggestion, helpful guidance, and encouragement in the execution of this project work.

We would like to thank our course mentor for guiding us in making the Project work successful.

TABLE OF CONTENTS

1. Abstract

Cancer is the most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Data mining is a powerful technique to help the people in their health, Scientific and Engineering. Those techniques are extracting the hidden information from the large databases which helps to find the relationships and patterns from the data. This proposal is used to develop a software based Self Organizing Map (SOM) structure which is used to discover the hidden patterns in the lung disorder CT images by using the data mining techniques. This approach starts by extracting the lung regions from the CT image using image processing techniques, including bit Image Slicing, Erosion and Weiner filtering. The bit plane slicing technique is used in the extraction process to convert the CT image into a binary image. Bit plane slicing technique is faster, data and user independent. So many algorithms were developed to detect lung cancer but they are not proved if the in dependent assumptions are taken into consideration. In the era of algorithms used for detecting lung cancer, a Software based SOM structure is not developed. This paper starts with visualizing the closed structure of the Lung regions and then the disorder dataset is process using SOM Toolbox to create a learned SOM using K-means clustering. By using this data mining technique with SOM, it has the advantages of robust to analysis and cost effective method. Data mining technique uses learning method to understand the data patterns. The SOM can be used within the data mining and exploratory data analysis process. Lung cancer is one of the leading causes of mortality in every country, affecting both men and women. Lung cancer has a low prognosis, resulting in a high death rate. The computing sector is fully automating it, and the medical industry is also automating itself with the aid of image recognition and data analytics. This paper endeavors to inspect accuracy ratio of three classifiers which is Support Vector Machine (SVM), KNearest Neighbor (KNN)and, Convolutional Neural Network (CNN) that classify lung cancer in early stage so that many lives can be saving. Basically, the informational indexes utilized as a part of this examination are taken from UCI datasets for patients affected by lung cancer. The principle point of this paper is to the execution investigation of the classification algorithms accuracy by WEKA Tool

2. INTRODUCTION

Lung cancer is the one of the leading cause of cancer deaths in both men and women. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. [1]. Treatment and diagnosis depend on the histological type of cancer, the stage (degree of spread), and the patient's performance status. Possible treatments include chemotherapy, and radio therapy , surgery Survival depends on stage, overall health, and other factors, but overall only fourteen percent of people diagnosed with lung cancer survive five years after the diagnosis. Symptoms that may suggest lung cancer include:

- Chronic coughing or change in regular coughing pattern,
- Dyspnea (shortness of breath with activity),
- Wheezing,
- Hemoptysis (coughing up blood),
- Chest pain or pain in the abdomen,
- Cachexia (weight loss, fatigue, and loss of appetite),
- Clubbing of the fingernails(uncommon),
- Dysphasia(difficulty swallowing),
- Pain in shoulder ,chest , arm,
- Dysphonia (hoarse voice),
- Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

Mortality and morbidity due to tobacco use is very high. Usually lung cancer develops in the wall or epithelium of the bronchial tree. It can start anywhere in the lungs and affect the part of the respiratory system. Lung cancer is mostly affects people between the ages of 55 and 65 and often it takes more years to develop . There are two types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer spreads and effects in different way, and is treated differently. If the cancer have both types feature, then it is called mixed small cell/large cell cancer. Non-small cell lung cancer (NSCLC) is most commonly appears than SCLC and it generally increases and spreads more slowly. Smoking is also related with SCLC and increases and spreads more quickly and form large tumors that can spread through the body. Generally they start many times in the bronchi near the center of the chest. Lung cancer death rate is depends on total amount of cigarette smoking. . Primary prevention activities include of cigarette Smoking, diet modification, and chemoprevention. Screening is a reasonably secondary prevention. Our

proposed method of finding the available Lung cancer patients is based on the systematic study of symptoms and risk factors. Some of generic indicators of the cancer diseases are Non-clinical symptoms and risk factors. Many carcinogens are found in the air we breathe, the water we drink and the food we eat. Pre-diagnosis should identify or narrow down the possibility of screening for lung cancer disease. Symptoms and risk factors (smoking, alcohol consumption, obesity and insulin resistance) had a significant effect in pre-diagnosis stage.[. The lung cancer diagnostic and prognostic problems are mainly in the scope of the broadly discussed classification problems. These problems have attracted by many researchers in computational intelligence, statistics, data mining, and fields. Cancer research is usually clinical and biological in nature, data driven statistical research has become a standard complement. Predicting the outcome of a disease is the most interesting and challenging tasks where to develop data mining applications. The usage of computers powered with automated tools, large volumes of medical data have been collected and made available to the medical research groups. As a result, Knowledge Discovery in Data bases , which may includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit lung patterns and relationships among large number of variables, and they able to predict the outcome of a disease using the historical cases stored within datasets. The objective of this paper is to summarize various review and technical articles on diagnosis of lung cancer. It gives overview of the current research being carried out on various lung cancer datasets using the data mining techniques and to enhance the lung cancer diagnosis.

Primary prevention activities include of cigarette Smoking, diet modification, and chemoprevention. Screening is a reasonably secondary prevention. Our proposed method of finding the available Lung cancer patients is based on the systematic study of symptoms and risk factors. Some of generic indicators of the cancer diseases are Non-clinical symptoms and risk factors. Many carcinogens are found in the air we breathe, the water we drink and the food we eat. Pre-diagnosis should identify or narrow down the possibility of screening for lung cancer disease. Symptoms and risk factors (smoking, alcohol consumption, obesity and insulin resistance) had a significant effect in pre-diagnosis stage.[4]. The lung cancer diagnostic and prognostic problems are mainly in the scope of the broadly discussed classification problems. These problems have attracted by many researchers in computational intelligence, statistics, data mining, and fields. Cancer research is usually clinical and biological in nature, data driven statistical research has become a standard complement. Predicting the outcome of a disease is the most interesting and challenging tasks where to

develop data mining applications. The usage of computers powered with automated tools, large volumes of medical data have been collected and made available to the medical research groups. As a result, Knowledge Discovery in Data bases , which may includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit lung patterns and relationships among large number of variables, and they able to predict the outcome of a disease using the historical cases stored within datasets. The objective of this paper is to summarize various review and technical articles on diagnosis of lung cancer. It gives overview of the current research being carried out on various lung cancer datasets using the data mining techniques and to enhance the lung cancer diagnosis.

The purpose of the data pre-processing is to simplify the training and testing process by appropriately transforming the entire dataset. This is done by bringing outliers and scaling properties into an equivalent range. Pre-processing is an important stage before training machine-learning models since pre-processing applications allow machine learning to deliver better results . The most used pre-processing applications are normalization, dimensional reduction of data, and feature selection. Normalization and dimensional reduction of data provide efficient, quick, and effective classifications through input and output of the prediction model in a single order in machine learning .

The Minimum–Maximum (min–max) normalization technique executes a linear transformation on the data. The minimum is the smallest value that data can accept and the maximum is the largest value that the data can receive. The data are usually in the range of 0–1 . The *Z*-score normalization method uses the average and standard deviation values of the data under consideration. It is typically utilized to perform the operations with the data at normal intervals with normalized values and to obtain more meaningful and easily interpreted results

Principal component analysis (PCA) is a transformation technique that reduces the size of the *p*-dimensional dataset that contains associated variables to a lower-dimensional space that contains uncorrelated variables. It maintains the variability of the dataset as much as possible. The variables obtained by transformation are called basic components of the original variables. The first principal component captures the maximum variance in the dataset, and the other components capture the remaining variance in descending order . The goal of the linear discriminant analysis (LDA) is to avoid over-alignment and also to project a dataset into a lower-dimensional space with good class separability to reduce computational costs. In addition, LDA is used to maximize the distance between classes. There is no class concept in PCA because it only reduces the feature size by maximizing the distance between data points. Likewise, LDA reduces the size of the dataset by maximizing the difference between classes. LDA is used for classification problems while PCA is used for clustering problems. LDA approximates the distance of the projections on the geometric plane of the observations in the same classes in an educational sample. It can be used to create more classifying models by normalizing the data in the training sample .

Feature selection is applied to reduce the number of features in many applications where the data have hundreds or thousands of properties. The main idea is to find globally the least reduction or, in other words, the smallest set of features that represent the most important characteristics of the original set of features .

The choice of methods for the machine-learning prediction system is important because there are many machine-learning algorithms used in practice for particular purposes. For

instance, random forest (RF) works with the logic of increasing the accuracy of results by deriving multiple decision trees while k-nearest neighbors (k-NN) uses similarities to find neighbors when classifying by majority vote. Naïve Bayes (NB) maintains the most appropriate classification by preserving the dependence of the qualifications on a particular class. Logistic regression (LR) finds the dependent and independent relationships between the variables affected by the dependent variables. Decision tree (DT) is the preferred learning method with a created tree structure since it is faster, easier to interpret, and is more effective. Support vector machines (SVMs) work with hyperplanes to separate data classification into a multidimensional space . k-NN has been reported to give the best results for machine-learning algorithms applied to the histopathological classification of non-small cell carcinomas. The DT algorithm was reported to give the least favorable results .

This study aims to develop predictive models to diagnose lung cancer disease based on a customized machine-learning framework. This approach involves examining the different degrees of success of these models and analyzes their generally valid classification performances according to measurement metrics. In this context, this study consists of three modules. The first module is based on the application of the data pre-processing techniques (dimensionality reduction methods, normalization techniques, and feature selection methods) to the lung cancer dataset (LCDS), which is taken from the Machine Learning Repository website of the University of California, Irvine (UCI). The second module focuses on the demonstration and discussion of the performance of the machine-learning algorithms (RF, k-NN, NB, LR, DT, and SVMs). The third module looks at the results of all the performance measurement metrics and performs validation analysis. The aforementioned methods are widely used in diagnosis and analysis to make decisions in different areas of medicine and research.

3. LITERATURE FOR LUNG CANCER

The approach that is followed for the prediction technique is depend on systematic study of the statistical factors, symptoms and associated with Lung cancer. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Initially parameters for the prediagnosis are collected by interacting with the pathological, clinical and medical oncologists (main experts).

A. Statistical Incidence Factors:

i. Age-adjusted rate (ARR)

ii. Area-related incidence chance

iii. Crude incidence rate

iv. Primary histology

B. Lung cancer symptoms:

The following are the generic lung cancer symptoms [7].

i. A cough that does not go away and gets worse over time
ii. Coughing up blood (heamoptysis) or bloody mucus.
iii. Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
iv. Weight loss and loss of appetite a. Wheezing b. Increase in volume of sputum
v. Fatigue and weakness
vi. Repeated problems with pneumonia or bronchitis
vii. Repeated respiratory infections, such as bronchitis or pneumonia
viii. Fatigue and weakness Shortness of breath
ix. New onset of wheezing
x. Swelling of the neck and face a. Clubbing of the fingers and toes. The nails appear to bulge out more than normal.
xi. Paraneo plastic syndromes which are caused by biologically active substances that are secreted by the tumor. i. Fever ii. Hoarseness of voice
xii. Loss of appetite xiii. Puffiness of face xiv. Nausea and vomiting

C. Lung cancer risk factors:

1. Smoking:
    a. Beedi
    b. Cigaratte
    c. Hukka
2. Second-hand smoke
3. Radon exposure
4. High dose of ionizing radiation
5. Occupational exposure to mustard gas chloro methyl ether, inorganic arsenic, chromium, nickel, radon asbestos
6. . Air pollution

4. MACHINE LEARNING

Machine learning is a subfield of Artificial Intelligence . Machine Learning is also used for complex data classification and decision making . In general, the implementation of algorithms aids the machine's learning. Machine learning gives systems the opportunity to learn automatically and improve over time without being directly configured. The implementation of algorithms aids the computer in learning and making the required decisions . Machine Learning strategies and activities are narrowly divided into three categories:

a. Supervised learning

Supervised learning is the machine learning task of a learning a function that maps an input to an output based on example of input-output pairs. It infers a function from labeled training data consisting a set of training Examples. Machine learning, in its most simple form, employs programmed algorithms that learn and refine their functions by processing input data and making predictions within a reasonable range. These algorithms aim to be predictive more precisely by feeding fresh data. While there are several changes in the way machine learning algorithms are grouped. Two categories of issues: grouping problems and back-up problems, are well suited to supervised learning algorithms. The output variable usually takes on a limited number of discrete values.

b. Un-Supervised Learning

Un-Supervised learning is a type of algorithms that learns patterns of untagged data. It refers to the use of AI algorithms to identify patterns in data set containing data points that's are neither classified nor labeled. The machine is given some sample inputs, but no output is generated in the method of learning. Since there is no optimal value over here, categorization is used to ensure that the algorithm distinguishes between the datasets correctly. It is the difficulty of finding unknown structure in unidentified details . Although there are no testing sets or tests given to the respondent, there are no opportunities to reward a successful solution. Unlike supervised learning and reinforcement learning, unsupervised learning has no teacher, and produces results that are unrelated to prior experience. It is directly connected to density and statistics.

c. Reinforcement Learning

Reinforcement learning is an area of Machine Learning . It is about talking suitable action to maximize reward in a particular situation. Reinforcement Learning, this machine learning style comes from interacting with its surroundings Reinforcement

learning. A Reinforcement Learning manager learns from the meaning of tasks, and even by explicitly articulated instructions, and decides on previous behaviors by using new techniques. Since specific input/output data sets are not provided, this differs from traditional supervised learning. Instead, the focus is on the presentation, which entails striking a balance between discovery (of uncharted territory) and utilization (of existing data).

5. DATABASES USED

I used Jupyter Notebook as my working platform or IDE during this Project . And the Dataset was downloaded from Kaggle. Different libraries and modules has been used in the process. The data used in this work is a lung cancer dataset that was first released in and later made available in the UCI machine learning repository under the name "Lung Cancer Data set". This dataset was used to show the capability of the optimum discriminant plane in ill-posed situations. This dataset contains data on the pathological forms of lung cancer. It contains 59 elements.

6. Objective

The main objective of the project is to construct a program used for the predictions of Lung cancer using python machine learning .

7. Background

In the background of this project ,used libraries are given below

- **Numpy –** Numpy is a library for python programming language, adding support for large ,multi-dimensional arrays and matrices along with a large collections of high-level mathematical functions to operate on the arrays.

- **Pandas -** Pandas is defined as open-source library that provides high-performance data manipulation in Python. It provides numerous functions and methods that expedite the data analysis and preprocessing steps.

- **Matplotlib -** Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

- **Pyplot –** It is a collection of functions in the popular visualization package Matplotlib . its functions manipulate elements of a figure, such as creating a figure , creating a plotting area , plotting lines , adding plot labels etc.

- **Seaborn -** Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.

- **Scilkit-Learn -** Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

- **Scatter_Matrix - A** scatter matrix (pairs plot) dataset against compactly plots all the numeric variables we have in a each other one. In Python, this data visualization technique can be carried out with many libraries but if we are using Pandas to load the data, we can use the base scatter_matrix method to visualize the dataset.

- **Tensorflow –** Tensorflow is a free and open-source software library for machine learning . it can be used across a range of tasks but has a particular focus on training and inference of deep neural network .

- **Keras-** Keras is a high level API on Neural Network. Which is developed in python language and can run on Tensorflow, Theano and CNTK.

  ➢ It is very user friendly

  ➢ It is well documented.

  ➢ It is scalable.

  ➢ It runs on CPU and GPU.

  ➢ It allows very fast and easy prototyping.

-

8. DATA MINING

Data Mining is major study of the data and Data Mining tools and techniques are used for discovery patterns from the data set. The most importance of Data Mining is to find patterns mechanically with least user input and efforts. Data Mining is an essential tool able of usage decision building and for forecasting expectations trends of market. Data Mining tools and techniques can be effectively functional in different fields in different domains. Many Organizations now begin using Data Mining as an efficient tool, to contract with the aggressive surroundings for data analysis. By using Data Mining tools and techniques, different fields of business get advantage by simply assess various trends of market and to make rapid and efficient market trend analysis. Data mining is very popular helpful tool for the diagnosis of disease.

Techniques

Techniques that are used in Data Mining Classification are a classic data mining technique based on machine learning. Majorly classification is used to classify every item in a set of data into one of predefined set of classes or groups. Classification technique makes mathematical techniques such as decision trees, linear programming, neural network and statistics.

Clustering :

Clustering is a data mining technique that makes useful or helpful cluster of substance that have similar feature using mechanical technique. From classification, clustering technique

defines the classes and keeps objects in them, as in classification objects are assigned into predefined classes. For example prediction of heart disease by using clustering obtain cluster or state that list of patients have same risk factor. Funds this makes the split list of patients with high blood sugar and related risk factor and so on.

Association:

Association is one of the best data mining techniques. In association, a pattern is exposed based on a relationship of a particular item on other items of the same operation. For example, the association technique is used in finding heart disease prediction as it say to us the relationship of dissimilar attributes used for analysis and sort out the patient with all the risk factor which are prediction of disease.

Prediction:

Used for profit prediction, relationship among dependent and independent variables and relationship between independent variables. For example, prediction analysis technique can be helped in sale to predict profit for the future if consider sale is an independent variable, profit could be a dependent variable.

DATA MINING PROCESS

In the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining process applied. Generally there are two methods of data mining tasks, descriptive data mining task is that describe the general properties of the existing data, and predictive data mining task is that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative and multimedia. Data mining applications can exhibit different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis, classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects).

Data mining involves some of the following steps.

Problem definition:- The first step is to identify goals. Based on the defined goal, the correct series of tools to be applied to the data to build the corresponding behavioral model.

Data exploration:- If the quality of data is not suitable for an accurate model then the recommendations on future data collection and storage strategies can be made for this. For analysis, all data must need to be consolidated so that it can be treated consistently.

Data preparation: The purpose of this step is to clean and transform the data, so that missing and invalid values are treated ,and all known valid values are made consistent for the more robust analysis.

Modeling:- Based on the data and desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

Evaluation and Deployment:- Based on the results the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

9.Related Work

In this project I try to predict lung cancer using different algorithms . Predictor variable use in classifying lung cancer:

1. Age
2. Smokes
3. AreaQ
4. Alkhol

Step for Creating Project:

- First I loaded the data set after that import the data
- Creating data comparison between predict value using graphs
- Split the dataset and train the data
- Predict the dataset using different algorithms
- Calculate the accuracy for all algorithms
- Find the best model for Dataset

10. Model of Dataset:

1. **Logistic Regression:-** Logistic regression is a technique that can be applied to binary classification problems. This technique uses the logistic function or sigmoid function, which is an S-shaped curve that can assume any real value number and assign it to a value between 0 and 1, but never exactly in those limits. Thus, logistic regression models the probability of the default class (the probability that an input (X) belongs to the default class $(Y=1)$) $(P(X)=P(Y=1|X))$. In order to make the prediction of the probability, the logistic function is used, which allows us to obtain the log-odds or the probit. Thus, the model is a linear combination of the inputs, but that this linear combination relates to the log-odds of the default class. Started from make an instance of the model setting the default values. Specify the inverse of the regularization strength in 10. Trained the logistic regression model with the training data, and then applied such model to the test data.

2. **Support Vector Machine :-** SVMs (Support Vector Machine) have shown a rapid proliferation during the last years. The learning problem setting for SVMs corresponds to a some unknown and nonlinear dependency (mapping, function) $y = f(x)$ between some high-dimensional input vector x and scalar output y. It is noteworthy that there is no information on the joint probability functions, therefore, a free distribution learning must be carried out. The only information available is a training data set $D = {(x_i, y_i) ∈ X×Y }, i = 1$, l, where l stands for the number of the training data pairs and is therefore equal to the size of the training data set D, additionally, y_i is denoted as d_i, where d stands for a desired (target) value. Hence, SVMs belong to the supervised learning techniques. From the classification approach, the goal of SVM is to find a hyperplane in an N-dimensional space that clearly classifies the data points. Thus hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes.

3. **K-Nearest Neighbor classification:-** K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. K-Nearest neighbors is a technique that stores all available cases and classifies new cases based on a similarity measure e.g., distance functions. This technique is non-parametric since there are no assumptions for the distribution of underlying data and it is lazy since it does not need any training data point model generation. All the training data used in the test phase. This makes the training faster and the test phase slower and more costlier. In this technique, the number of neighbors k is usually an odd number .

4. **Decision Tree Classification:-** A decision tree is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule, and each leaf node represents the outcome. The decision tree analyzes a set of data to construct a set of rules or questions, which are used to predict a class, i.e., the goal of decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. In this sense the decision tree selects the

best attribute using to divide the records, converting that attribute into a decision node and dividing the data set into smaller subsets, to finally start the construction of the tree repeating this process recursively.

11. Software Tools:

- Jupyter Notebook

12. Source code

- **First import the libraries-**

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import accuracy_score,confusion_matrix
```

- **Load the Dataset-**

```
print("Dataset:")
dataset=pd.read_csv('Lung_Cancer_Dataset.csv')
print(len(dataset))
print(dataset.head())
```

Output-

```
Dataset:
59
     Name      Surname  Age  Smokes  AreaQ  Alkhol  Result
0    John         Wick   35       3      5       4       1
1    John  Constantine   27      20      2       5       1
2  Camela     Anderson   30       0      5       2       0
3    Alex       Telles   28       0      8       1       0
4   Diego     Maradona   68       4      5       6       1
```

- **Checking dataset has null values-**

```
sns.heatmap(dataset.isnull())
```

Output-

```
<AxesSubplot:>
```

- **Understanding of Data-**

  Print the dataset Result-
  dataset['Result'].value_counts()

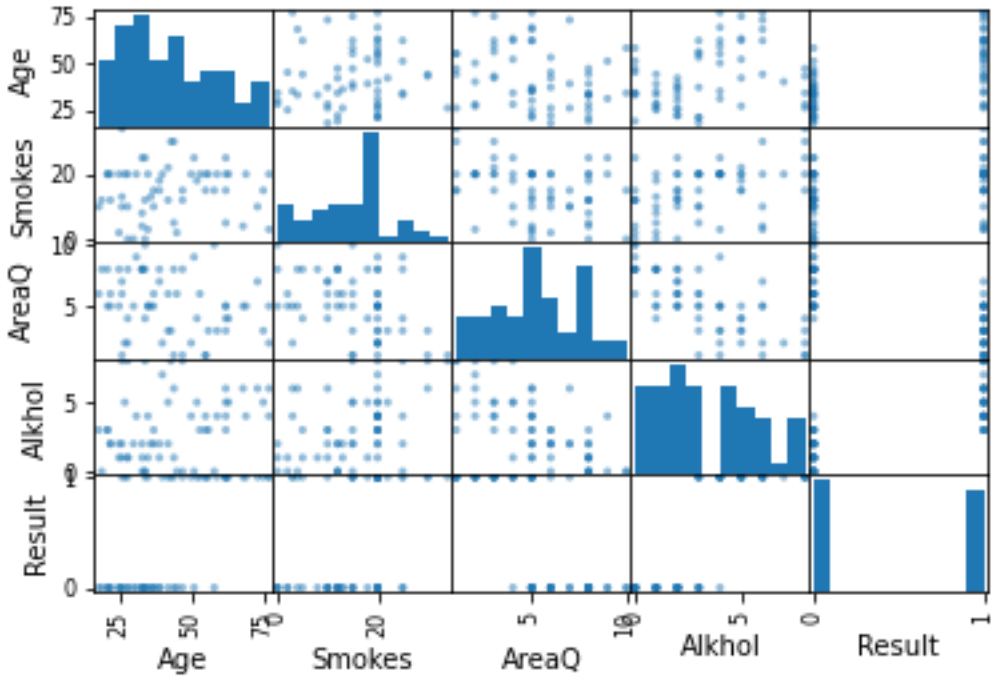
  output-

```
0    31
1    28
Name: Result, dtype: int64
```

- Using Scatter_matrix understand the dataset-

  from pandas.plotting import scatter_matrix
  from matplotlib import pyplot
  scatter_matrix(dataset)
  pyplot.show

  Output-

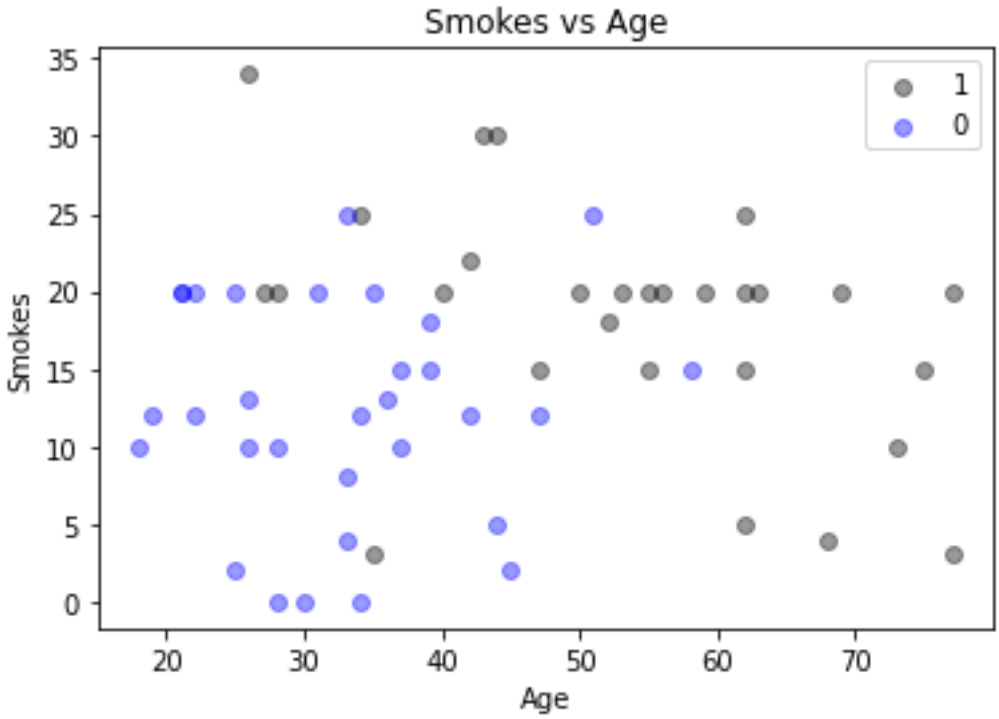


- Taking two variable and store the label of dataset-
  A=dataset[dataset.Result==1]
  B=dataset[dataset.Result==0]

- **Data Analysis Smoke vs Age**

  plt.scatter(A.Age ,A.Smokes , color="Black", label="1", alpha=0.4)
  plt.scatter(B.Age ,B.Smokes, color="Blue", label="0",alpha=0.4)
  plt.xlabel("Age")
  plt.ylabel("Smokes")
  plt.legend()
  plt.title("Smokes vs Age")
  plt.show()

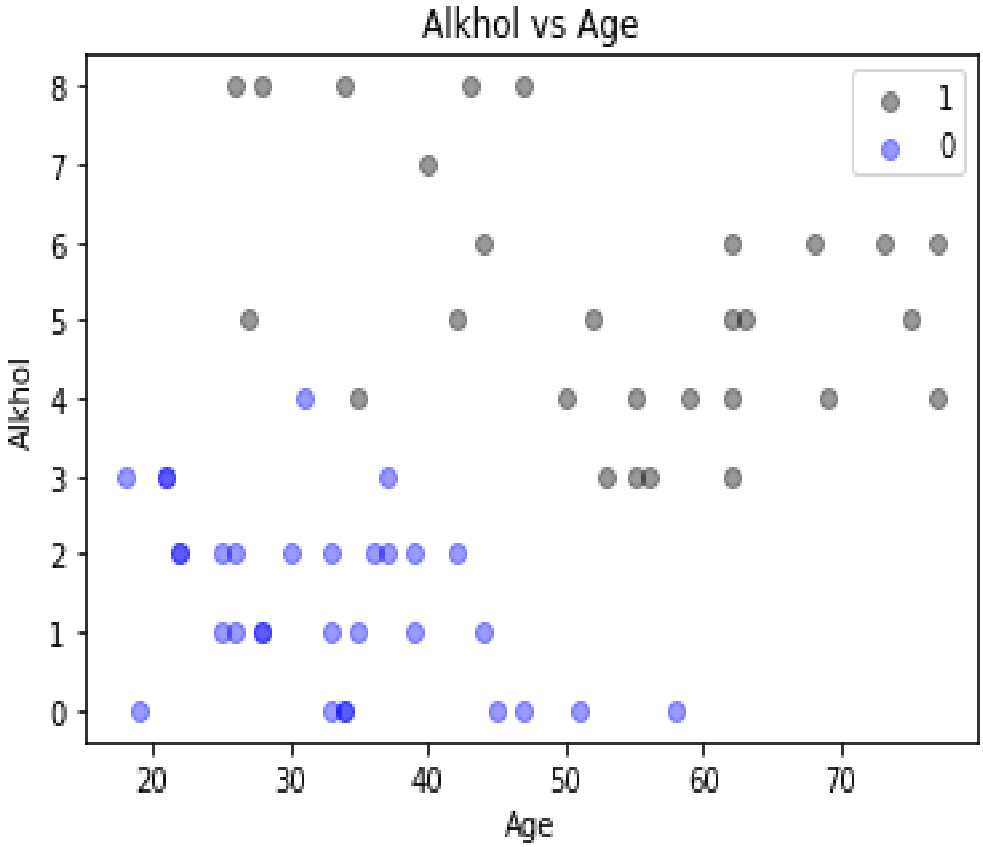
  Output-



- **Data Analysis Alkhol vs Age-**

```
plt.scatter(A.Age, A.Alkhol, color="Black",label="1",alpha=0.4)
plt.scatter(B.Age, B.Alkhol, color="Blue",label="0",alpha=0.4)
plt.xlabel("Age")
plt.ylabel("Alkhol")
plt.legend()
plt.title("Alkhol vs Age")
plt.show()
```

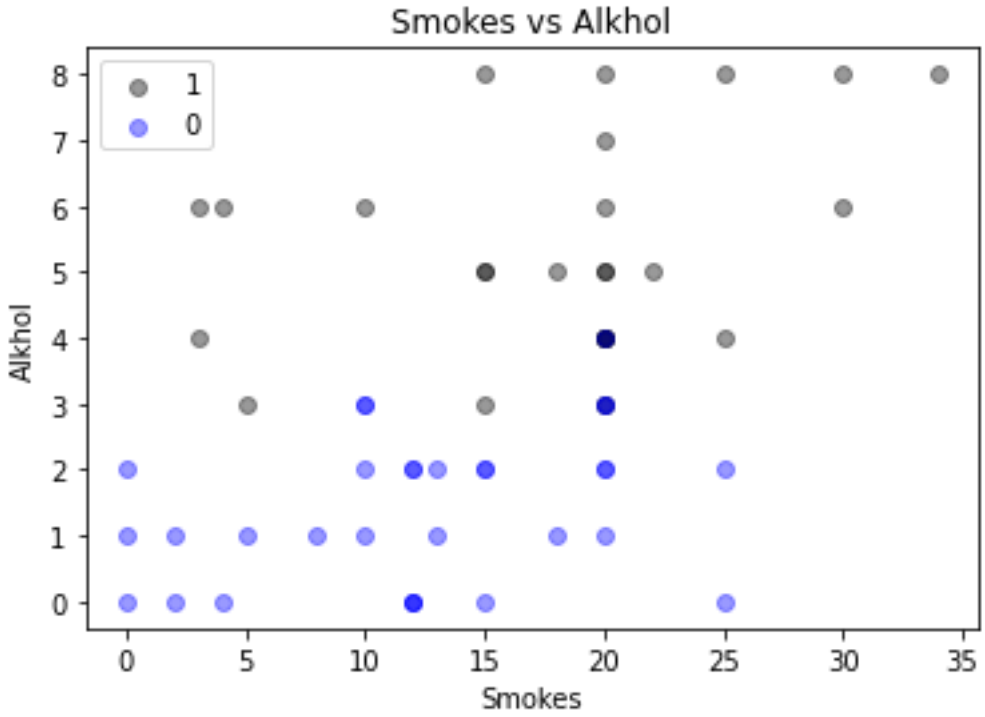
output-



- **Data analysis Smokes vs Alkhol**

```
plt.scatter(A.Smokes, A.Alkhol, color="Black",label="1",alpha=0.4)
plt.scatter(B.Smokes, B.Alkhol, color="Blue",label="0",alpha=0.4)
plt.xlabel("Smokes")
plt.ylabel("Alkhol")
plt.legend()
plt.title("Smokes vs Alkhol")
plt.show()
```

Output-



- **Split the dataset and train the data-**

  from sklearn.model_selection import train_test_split
  from sklearn.preprocessing import StandardScaler

  #split dataset
  x=dataset.iloc[:,3:5]
  y=dataset.iloc[:,6]
  x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=0,test_size=0.2
)

  #feature Scaling
  sc_x=StandardScaler()
  x_train=sc_x.fit_transform(x_train)
  x_test=sc_x.transform(x_test)

  Using the Different algorithms for Predict the Data-
  Model Implementation- Model is implemented using four types of algorithms.

- **Logistic Regression-**

  it predicts a binary outcome ,such as yes or no, based on prior observations of a data set.

  from sklearn.linear_model import LogisticRegression

```
# We defining the model
logreg = LogisticRegression()

# We train the model

logreg.fit(x_train, y_train)

# We predict target values

y_predict1 = logreg.predict(x_test)

#Evaluate the dataset using confusion matrix

from sklearn.metrics import confusion_matrix
logreg_cm=confusion_matrix(y_test,y_predict1)
print("Confusion Matrix:")
print(logreg_cm)

output-

Confusion Matrix:
[[8 0]
[1 3]]

#calculate the accuracy

from sklearn.model_selection import cross_val_score
accuracies=cross_val_score(estimator=logreg,X=x_train,y=y_train,cv=10)
print("accuracy is in percentage",format(accuracies.mean()*100))
```

accuracy is in percentage 82.0

- **Support Vector Machine-**

  a supervised machine learning algorithm that can be employed for both classifi cation and regression purposes.

```
from sklearn.svm import SVC

#we define the model
sv=SVC(kernel='linear',random_state=0)

#train the model
sv.fit(x_train,y_train)

#predict the model
y_predict2 = sv.predict(x_test)
```

```
# Evaluate the dataset using confusion matrix
from sklearn.metrics import confusion_matrix
sv_cm=confusion_matrix(y_test,y_predict2)
print("Confusion Matrix:")
print(sv_cm)
```

Output-

```
Confusion Matrix:
[[8 0]
[1 3]]
```

```
#calculate the accuracy

from sklearn.model_selection import cross_val_score
accuraciess=cross_val_score(estimator=sv,X=x_train,y=y_train,cv=10)
print("accuracy is in percentage",format(accuraciess.mean()*100))
```

accuracy is in percentage 81.5

- **K-Nearest Neighbor classification-**

supervised machine learning algorithm that can be used to solve both classification and regression problem.

```
from sklearn.neighbors import KNeighborsClassifier

kncc=KNeighborsClassifier(n_neighbors=5,n_jobs=-1)

kncc.fit(x_train, y_train)

y_predict3=kncc.predict(x_test)

#Evaluate the model

from sklearn.metrics import confusion_matrix

kncc_cm=confusion_matrix(y_test,y_predict3)

print("Confusion Matrix:")

print(kncc_cm)
```

Output-

```
Confusion Matrix:
[[7 1]
[1 3]]
```

```
#calculate the accuracy
```

```
from sklearn.model_selection import cross_val_score
accuracies_kncc=cross_val_score(estimator=kncc,X=x_train,y=y_train,cv=10)
print("accuracy is in percentage",format(accuracies_kncc.mean()*100))
```

accuracy is in percentage 86.5

- ## Decision tree classification-

The decision tree analyzes a set of data to construct a set of rules or questions, which are used to predict a class, i.e., the goal of decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

```
from sklearn.tree import DecisionTreeClassifier
dtc= DecisionTreeClassifier()
dtc.fit(x_train, y_train)
y_predict4 = dtc.predict(x_test)
#Evaluate the model
from sklearn.metrics import confusion_matrix
dtc_cm=confusion_matrix(y_test,y_predict4)
print("Confusion Matrix:")
print(dtc_cm)
```

Output-

```
Confusion Matrix:
[[5 3]
 [0 4]]
```

```
#calculate the accuracy
from sklearn.model_selection import cross_val_score
accuracies_dtc=cross_val_score(estimator=dtc,X=x_train,y=y_train,cv=10)
print("accuracy is in percentage",format(accuracies_dtc.mean()*100))
```

accuracy is in percentage 71.0

Accuracies of all classification model overview-

Logistic regression :82%

support vector machine:81%

K-nearest Neighbor classification:86%

Decision tree classification:75%

The best model is K-Nearest Neighbor with 86% Accuracy

13.Related Concepts-

Nodule- A nodule is a small round oval shaped growth in the Lung. It may also be called a spot on the lungs or a coin lesion. Nodules are smaller than 3 centimeters in diameter. if the growth is larger than that it is called a pulmonary mass and is more likely to represent cancer . it has two types namely malignant and benign. Malignant are cancerous. Unfortunately no apparent symptoms are associated with its presence and they can only be detected with computed tomography or traditional x-rays.

Nodules have different classes and each have a different impact on their growth rates, intensities, sizes, detections and cure.

Classification of Nodules:

Lung nodules can be distinguished in solid nodules and sub solid nodules. Sub solid nodules can be further classified as nonsolid nodules and part solid nodules. This classification is significant because different nodules require different approaches for their detection, measurement & management.

Lung Cancer debrief-

In simple words cancer is abnormal growth of cells, and ultimately to the stoppage of the essential cellular functions of the organism. These cells are generally called ‘tumor cells’ and they often clump together into lumps to form ‘tumors’. They also carry the potential to invade to other parts of the body of the organism and have detrimental effects there.

Lung cancer can be divided into four stages depending on severity as following-

- Stage 1: Cancer is found in the lung only.
- Stage 2: Cancer is grown to nearby lymph nodes.
- Stage 3: Cancer is in the lung and lymph nodes are grown in the middle of the chest.
- Stage 3A: Cancer is found in lymph nodes only on the same side of the chest where it first started.

● Stage 3B: Cancer has spread to lymph nodes on the opposite side of the chest or to lymph nodes above the collarbone.
● Stage 4: Cancer has spread to both lungs, into the area around the lungs, or to distant organs

Early symptoms may include-

● Bad painful cough

● Cough with phlegm or blood

● Heavy chest pain while deep breathing, laughing, coughing or any other muscular activities near chest

● shortness of breath

● weakness and fatigue

● loss of appetite and weight loss

Causes

About ninety percent of lung cancer cases involve smoking. Smoking causes destruction of lung tissues, lung can repair them but heavy smoking makes lung stopping it's natural behaviour. A radioactive gas Radon, is the second leading cause, according to the American Lung Association. Breathing in other hazardous substances can also cause lung cancer if it happens over a long period of time. A type of lung cancer called mesothelioma is almost always caused by exposure to asbestos.

Other substances that can cause lung cancer are:

● nickel

● petroleum products

● uranium

● arsenic

● Cadmium

● Chromium

Inherited genetic mutations may be the reason to develop lung cancer, especially if you smoke or are exposed to other carcinogens. But above all there may be no specific reason for lung cancer.

Risks Factor-

Above all the biggest risk factor is smoking. Inhaling toxic substances increases risk of getting lung cancer. Secondhand smoke is also a major risk factor. Other rish facors may include family health history, previous therapy or health history etc.

Diagnosis-

Imaging tests:- An abnormal lump can be seen on X-ray. MRI, CT, and PET scans. These scans produce more detail and find smaller lesions.

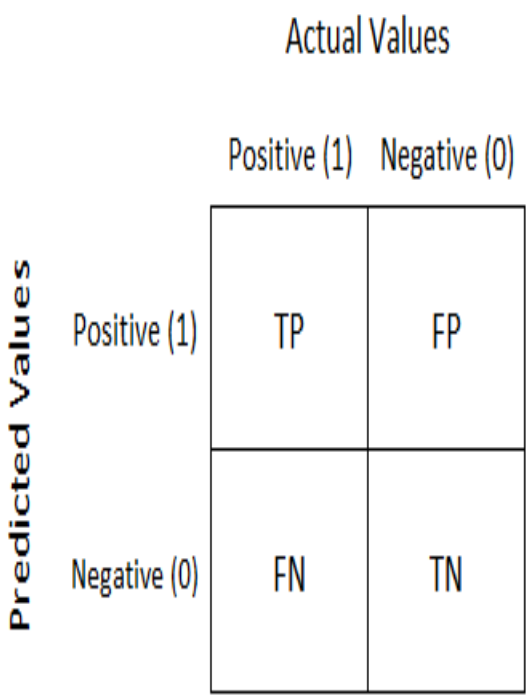
Sputum cytology:- Microscopic examination of phlegm of cough can determine if cancer cells are present.

Treatment-

- Self-care
- Quitting smoking
- Medications
- Chemotherapy and Targeted therapy
- Surgery
- Pulmonary lobectomy and Video-Assisted thoracoscopic surgery
- Medical procedure
- Thoracotomy and Radiation therapy
- Supportive care
- Palliative care
- Specialists

Confusion Matrix- The Confusion Matrix is a deep learning visual assessment method. T he prediction class results are represented in the columns of a Confusion Matrix, whereas the real class results are represented in the rows [54]. This matrix includes all the raw data regard ing a classification model's assumptions on a specified data collection. To determine how acc urate a model is. It's a square matrix with the rows representing the instances' real class and th e columns representing their expected class. The confusion matrix is a 2 x 2 matrix that report s the number of true positives (T P), true negatives (T N), false positives (FP), and false negat ives (F N) when dealing with a binary Precision, recall, and F-measure, which are commonly utilized in the text mining and machine learning communities, were used to evaluate the algor ithms. True positive (TP – objects correctly labeled as belonging to the class), false positive (FP – items falsely labeled as belonging to a certain class), false negative (FN – items incorrec tly labeled as not belonging to a certain class), and true negative (TN – items incorrectly label ed as not belonging to a certain class) are the four types of classified items (TN - items correc tly labelled as not belonging to a certain class). Recall is determined using the following form ula given the amount of true positives and false negatives.

The confusion matrix was used to evaluate the accuracy of each classifier. The experimental results show that using five attributes from an SVM classifier produces the best classification mission.



Future Selection- Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection. Each machine learning process depends on feature engineering, which mainly contains two processes; which are Feature Selection and Feature Extraction. Although feature selection and extraction processes may have the same objective, both are completely different from each other. The main difference between them is that feature selection is about selecting the subset of the original feature set, whereas feature extraction creates new features. Feature selection is a way of reducing the input variable for the model by using only relevant data in order to reduce overfitting in the model.

Need for future Selection- Before implementing any technique, it is really important to understand, need for the technique and so for the Feature Selection. As we know, in machine learning, it is necessary to provide a pre-processed and good input dataset in order to get better outcomes. We collect a huge amount of data to train our model and help it to learn better. Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data. Moreover, the huge amount of data also slows down the training process of the model, and with noise and irrelevant data, the model may not predict and perform well. So, it is very necessary to remove such noises and less-important data from the dataset and to do this, and Feature selection techniques are used.

Selecting the best features helps the model to perform well. For example, Suppose we want to create a model that automatically decides which car should be crushed for a spare part, and to do this, we have a dataset. This dataset contains a Model of the car, Year, Owner's name,

Miles. So, in this dataset, the name of the owner does not contribute to the model performance as it does not decide if the car should be crushed or not, so we can remove this column and select the rest of the features(column) for the model building.

Below are some benefits of using feature selection in machine learning:

- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- It reduces the training time.
- It reduces overfitting hence enhance the generalization.

Clustering- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

K-means clustering algorithm - It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.

There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

Self Organizing maps-

The self-organizing map (SOM) is an excellent tool in exploratory phase of the data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively used to visualize and explore properties of the data. When number of SOM units is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped, i.e. clustered. In this paper, we analyze different approaches to clustering of the SOM are considered. The two-stage procedure-first using SOM to produce the proto types then that are clustered in the second stage, that is found to perform well when compared with direct clustering of the data and to reduce the computation time.

Risk Models-

There have been a number of lung cancer risk models developed and validated that one may consider to be a form of CADx tool . Typically based on logistic regression, such tools aim to provide an overall risk of the patient having cancer based on patient meta-data such as age, sex and smoking history and nodule characteristics such as nodule size, morphology and growth, if a previous CT was available.

Although such tools currently require manual entry by the user, they do produce an objective lung cancer risk score which may be used in the decision-making process. However, despite their attraction and good performance, their adoption and performance as part of decision making has not been studied. The British Thoracic Society (BTS) guidelines on the management of incidentally detected pulmonary nodules recommends the use of the Brock model . Anecdotally, many physicians report using them for patient communication only and feel that such models do not add a great deal to their clinical expertise. More specifically, questions remain as to the utility of such models when the patient population is different to that of the training data. It is clear, that for such models to be clinically useful, knowledge of the training data used is critical, and this also will determine the clinical scenarios in which they may be used. There are clearly significant differences in the pre-test probabilities of a nodule being malignant in different patient groups. For instance, patients with a current or prior history of malignancy are at significantly different risk of nodule malignancy than non-smokers with no significant prior history.

From a technical perspective, such models have a number of limitations. Foremost is the reliance on human interpretation of input variables such as nodule size, morphology and even the reliance on the patient's own estimate of factors such as smoking history. For example, under the Brock model, a 1mm increase in the reported size of a 5 mm spiculated solid nodule in a 50-year-old female almost doubles its risk, from 0.98% to 1.89%. However, inter-radiologist variability in reporting nodule size is typically greater than this . Moreover, inter-reader variability in reporting morphology and nodule type is common even amongst experienced thoracic radiologists .Some recent work to address this has been proposed by Ciompi *et al.* where an automated system for the classification on nodules into solid, non-solid, part-solid, calcified, perifissural and spiculated types was proposed. Overall classification accuracy is reported to be within the inter-radiologist variability at 79.5% but this varies between 86% for solid and calcified nodules down to 43% for spiculated nodule classification. Of course, since the ground-truth classifications were provided by radiologist opinion, the performance at validation cannot be expected to improve on that. As the authors point out, the nodule types are radiologist developed concepts that, while useful for clinical purposes, lack a precise definition. The impact of the system's output as an input to the Brock model was not reported and ultimately this approach should be judged on its ability to improve malignancy prediction.

Radiomics-

The term Radiomics refers to the automatic extraction of quantitative features from medical images and has been the subject of a great deal of investigation with applications including automated lesion classification, response assessment and therapy planning. Fundamentally,

the Radiomic approach aims to turn image voxels into a set of numbers that characterize the biological property of interest such as lesion malignancy, tumour grade or therapy response.

Although research into, what are termed, Radiomics methods has seen an explosion in the last decade, the technical methods that it builds on have a very long history in the fields of computer vision and medical image understanding in the area of texture analysis. Indeed, many of the so-called Radiomic features are based on techniques that were first proposed in the 1970s for the classification of textured images and have been largely superseded in the computer vision literature. Nevertheless, their application to medical image processing research has in some areas yielded some significant insights, in particular in how such quantitative features relate to tumour pheno- and genotypes. The idea that such advanced quantitative techniques may add to the qualitative clinical interpretation of radiologists is gaining momentum and is likely to move into mainstream clinical practice in the coming 5 to 10 years.

For a given application, the Radiomic approach proceeds in two phases—first a training or feature selection phase and then a second testing or application phase. The training phase typically proceeds as follows. First, a large set, typically some hundreds or thousands, of features are defined a-priori. Next, the features are extracted from a large corpus of training data where the object of interest, say a tumour, has been delineated such that a computer algorithm can extract the quantitative features automatically. Finally, a step known as feature selection is applied that aims to select a smaller subset, e.g., some tens of such features that efficiently captures the imaging characteristics of the biological phenomena of interest. For example, in the case of nodule classification into benign and malignant we may pick the features, either individually or in combination that perform the best at this task on the training data.

In the testing phase, the Radiomics are applied to a particular patient's image, with the process being similar to the training phase but now the selected features are identified by the algorithm, extracted and then used to classify the patient.

Of course, both at training and testing steps, a classification algorithm will need to be defined to convert the Radiomics values into classifications. For small sets of individual features, we may simply use thresholds on the Radiomic features; however, for larger sets of features more sophisticated techniques from the field of machine learning, such as Support Vector Machines (SVMs) and Random Forests are typically used to yield better results. A very good review of Radiomics approaches applied to the classification of pulmonary nodules is provided in Wilson *et al*.

One criticism of some of the earlier Radiomics work is the lack of independent training and validation data . Indeed, it is not unusual to find very high classification rates being reported based on the training data whereas it is well established within the machine learning literature that such results may be subject to "overfitting"—the apparent excellent performance that cannot be replicated on unseen and independent datasets. In fact, one measure of the goodness of a well-trained classifier is the difference in performance between training sets and test sets. This phenomenon has led to a generally over-optimistic view of the performance with area under the curve (AUC) numbers reported in the high 80s and 90s range that cannot be replicated on independent data.

Deep Learning- Deep learning is a type of machine learning techniques that uses representation learning to categorize important features for classification problems [6]. The primary characteristic of deep learning is its compatibility with features, although it may also learn from data. So, to learn complex features a deep learning integrates the simple features that have learned from data. Deep learning is accomplished using multiple-layer artificial neural networks, such as the Deep Neural Network (DNN), Convolutional Neural Network (CNN), and the Recurrent Neural Network (RNN).

Convolutional Neural Network- CNN as a supervised deep learning tool, CNN is an excellent choice. This algorithm is suitable for multi-class classification and binary classification (for example, predicting whether or not a diagnostic picture contains a malignant tumor) [48][49]. CNNs are often used to solve a wide range of pattern and image recognition issues. This deep learning approach is effective and appropriate for visual data because of three key characteristics. To begin with, local receptive fields are perfectly matched to the image data specificity of being correlated geographically but uncorrelated globally. Second, since the convolution is applied to the entire image, mutual weights allow for significant parameter reduction without affecting image processing. Finally, a grid-structured image allows for data pooling operations that reduce data complexity without sacrificing valuable information.

14. Future Scope

- The lung cancer predictions using the machine learning algorithm can be used in early diagnosis of cancer in individuals. This can be very helpful for doctors , radiologist, for giving a better result for the patients who consult them. This technique can be used in curring the individuals and can also control the occurrence of lung cancer and can save millions of life . Machine learning algorithms can also be used for more such improvement in the health care and also the other sectors. most of the studies that have been proposed the last years and focus on the development of predictive model using supervised ml methods and classification algorithms aiming to predict valid disease.

15. Conclusion:-

Lung cancer is one of the most dangerous diseases and the most common cause of death, the severity of the disease lies in the difficulty of diagnosing it in the early stages. This paper tries to endeavor to investigate of four classifiers to find the best classifier could classify lung cancer in early stage. A prototype lung cancer disease prediction system is developed using different classifier algorithms. For example Logistic Regression, Support Vector Machine , K-Neighbour Classification algorithms and Decision tree algorithms. The system extracts hidden knowledge from a historical lung cancer disease database. Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining.

16 . BIBLIOGRAPHY

1. https://www.researchgate.net/
2. https://www.kaggle.com