A Project Report On

“**Lung Cancer Detection**”

Submitted in partial fulfilment of the

Requirement for the Award of degree of

**Bachelor of Technology**

of

Jamia Hamdard University



**Compiled by**:

Suhail Rizvi

(2020-310-221)

**Under the guidance of**:

Mr. Arvind Kumar

Manager Programming

GEOPIC, ONGC, Dehradun

****

# DECLARATION

I hereby declare that the work presented in this Project report entitled “**Lung Cancer Detection”** in partial fulfilment of the requirement for the Award of degree of **Bachelor of Technology** submitted in the department of Information Technology, Dehradun Institute of Technology, Dehradun is a record of my work carried out, in **Online Mode**, during the Summer Training from 21st June, 2023 to 21st July 2023, under the guidance of Mr. Arvind Kumar, Manager Programming, GEOPIC, ONGC, Dehradun.

**Signature of the Student**

Suhail Rizvi

# CERTIFICATE

This is to certify that Mr. Suhail Rizvi, a student of B.Tech. (CSE) of Jamia Hamdard University, Delhi has done his Summer Training, in **Online mode**, at **GEOPIC, ONGC, Dehradun** from 21st June,2023 to 21st July,2023. The project work entitled “**Lung Cancer Detection”** embodies the work done by him during the above-mentioned Summer training period.

**Signature of Project Guide Signature of Training Coordinator**

**(Mr. Arvind Kumar) (Mr. P R Meena)**

Manager Programming G.M (Programming)

**ACKNOWLEDGEMENT**

I deeply express my sincere thanks to my Internal guide Mr.Arvind Kumar (Manager Programming) who assisted me throughout this project. I thank him for providing me the right track for the project whenever I needed it. This project has been a success because of his guidance.

I’ll also express my sincere thanks to Mr. P R Meena my training co-ordinator, who guided me through the project for his valuable suggestions and guidance for completing the project.

Suhail Rizvi

**Contents**

|  |  |
| --- | --- |
| **TITLE** | **PAGE NO.** |
| 1. Introduction to Company ONGC | 6 |
| 1. Abstract | 7 |
| 1. Objective of the project | 8 |
| 1. Introduction | 9 |
| 1. Software Requirements Specifications | 10-11 |
| 1. Code and Snapshots of output screens | 12-15 |
| 1. Conclusion | 16 |

# 

**1.Introduction to Company-ONGC**

Maharatna ONGC is the largest crude oil and natural gas Company in India, contributing around 70 per cent to Indian domestic production. Crude oil is the raw material used by downstream companies like IOC, BPCL, and HPCL to produce petroleum products like Petrol, Diesel, Kerosene, Naphtha, and Cooking GasLPG. This largest natural gas company ranks 11th among global energy majors (Platts). It is the only public sector Indian company to feature in Fortune’s ‘Most Admired Energy Companies’ list. ONGC ranks 18th in ‘Oil and Gas operations’ and 183rd overall in Forbes Global 2000. Acclaimed for its Corporate Governance practices, Transparency International has ranked ONGC 26th among the biggest publicly traded global giants. It is most valued and largest E&P Company in the world, and one of the highest profit-making and dividend-paying enterprise

**Vision**

To be global leader in integrated energy business through sustainable growth, knowledge excellence and exemplary governance practices.

**Mission**

**World Class**

▪ Dedicated to excellence by leveraging competitive advantages in R&D and technology with involved people.

▪ Imbibe high standards of business ethics and organizational values.

▪ Abiding commitment to safety, health and environment to enrich quality of community life.

▪ Foster a culture of trust, openness and mutual concern to make working a stimulating and challenging experience for our people.

▪ Strive for customer delight through quality products and services

**Integrated in Energy Business**

▪ Focus on domestic and international oil and gas exploration and production business opportunities.

▪ Provide value linkages in other sectors of energy business.

▪ Create growth opportunities and maximize shareholder value.

**Dominant Indian Leadership**

▪ Retain dominant position in Indian petroleum sector and enhance India's energy availability.

**2. Abstract**

There are many assets related and incorporated by any multinational company for various works. This sometimes leads to tedious amount of work to monitor and manage. Although a relief is such that all these assets are connected to each other via a common network which makes it easier to track. The application designed by us aims to monitor the assets employed by a company using an IP range to extract crucial details about every node in a network. This is done by taking an IP range as an input and then producing each node within this range as an output. A user can choose to view various details about any system.

**3. OBJECTIVE**

Lung Cancer is considered as the deadliest cancer worldwide. For this reason , many countries are developing strategies for the early diagnosis of lung cancer. In this project the objective is to give the best result accuracy of lung cancer patients. To achieve this objective ,

we use KNeighbors classifiers and decision tree algorithms to classify the dataset and give the best accuracy result.

Using machine learning approaches for classifying whether a person is suffering from lung cancer or not , using the given dataset.

# 4.PROJECT INTRODUCTION

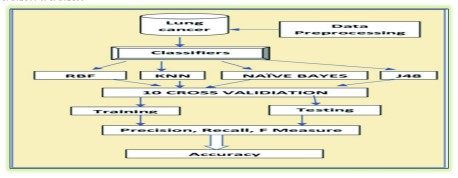
# Cancer is the disease in which cells in the body grow out of control. When cancer starts in the lungs it is called lung cancer. Lung cancer is the leading cause of cancer death and second most diagnosed cancer in both men and women in the United States. Cigarette Smoking is the number one cause of cancer. Lung cancer can also be caused by tobacco, breathing second-hand smoke being exposed to substances such as asbestos or radon at work. There are types of lung cancer and this cancer can be diagnosed by doctors with their procedure and to reduce the human efforts or human error for which we have developed a code in which we take the CT scan image and we define the properties and through the various algorithms we can able to detect the image is cancerous or not. In this world not only men but women also suffering from the same dangerous disease. After the detection, the life span of the patient suffering from lung cancer is very less. In this project I have used KNN & decision tree algorithms to predict the accuracy of cancer. In this project we use scikit-learn libraries like sklearn and pandas to predict and classify the dataset of lung cancer patients. Slicing the dataset and feature scaling options are used to trainthe dataset. After that we use a confusion matrix , f1 score and accuracy score to predict the accuracy of the result.

## 5. SOFTWARE REQUIREMENT SPECIFICATION

To perform KNN algorithm and decision tree algorithm we use scikit-learn library. In Background of this project , used libraries are given below.

1. Numpy- NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of arrays.
2. Pandas-Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.
3. Sklearn- Scikit-learn is a free software machine learning library for the Pythonprogramming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
4. Confusion Matrix- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.
5. KNeighbors Classifiers-The K in the name of this classifier represents the k nearestneighbors, where k is an integer value specified by the user. Hence as the name suggests, this classifier implements learning based on the k nearest neighbors. The Choice of the value of k is dependent on data.
6. Decision Tree- Decision Trees are a type of Supervised Machine Learning where the data is continuously split according t
7. A certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

# 



**6.PROGRAM CODE**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from pandas.plotting import scatter\_matrix

from matplotlib import pyplot

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import f1\_score

from sklearn.metrics import accuracy\_score

from sklearn import tree

from sklearn.metrics import confusion\_matrix

print("Dataset:")

dataset=pd.read\_csv("/content/lung\_cancer\_examples.csv")

print(len(dataset))

print(dataset.head())

scatter\_matrix(dataset)

pyplot.show()

A=dataset[dataset.Result==1]

B=dataset[dataset.Result==0]

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()

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()

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()

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)

sc\_x=StandardScaler()

x\_train=sc\_x.fit\_transform(x\_train)

x\_test=sc\_x.transform(x\_test)

print('--------------Using KNN Algorithm----------------')

import math

a=math.sqrt(len(y\_train))

print(a)

classifier=KNeighborsClassifier(n\_neighbors=5, p=2, metric='euclidean')

classifier.fit(x\_train, y\_train)

y\_pred=classifier.predict(x\_test)

print(y\_pred)

cm=confusion\_matrix(y\_test, y\_pred)

print("Confusion Matrix: ")

print(cm)

print("In Confusion Matrix:----- ")

print("Position 1.1 shows the patients that don't have cancer")

print("Position 1.2 shows the number of patients that have higher risk ofcancer")

print("Position 2.1 shows the incorrect value")

print("Position 2.2 shows the correct number of patients that have cancer")

print('F1 Score : ', (f1\_score(y\_test,y\_pred))\*100)

print('Accuracy : ',(accuracy\_score(y\_test, y\_pred))\*100)

print("----------Using Decsion Tree Algorithm----------")

c=tree.DecisionTreeClassifier()

c.fit(x\_train, y\_train)

accu\_train=np.sum(c.predict(x\_train)==y\_train)/float(y\_train.size)

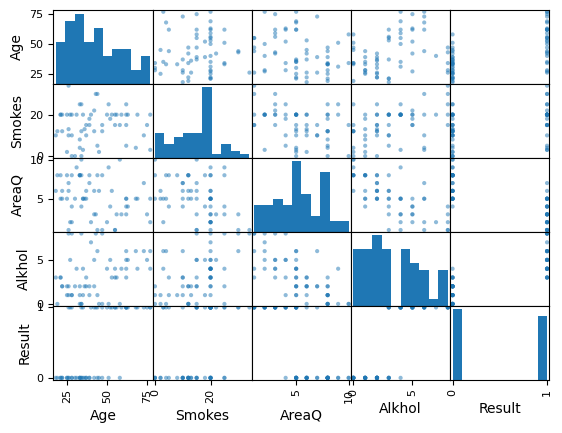
accu\_test=np.sum(c.predict(x\_test)==y\_test)/float(y\_test.size)

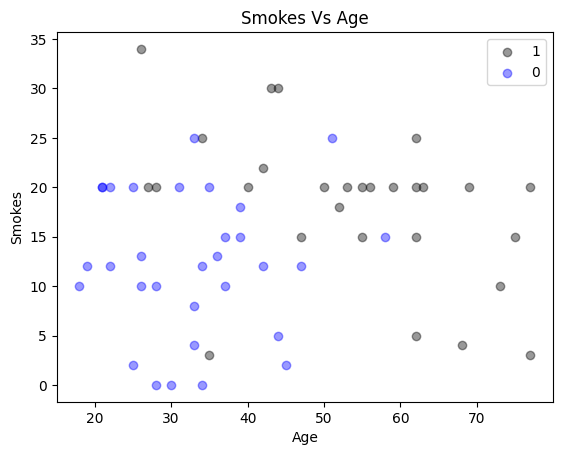
print('Classification accuracy on train', (accu\_train)\*100)

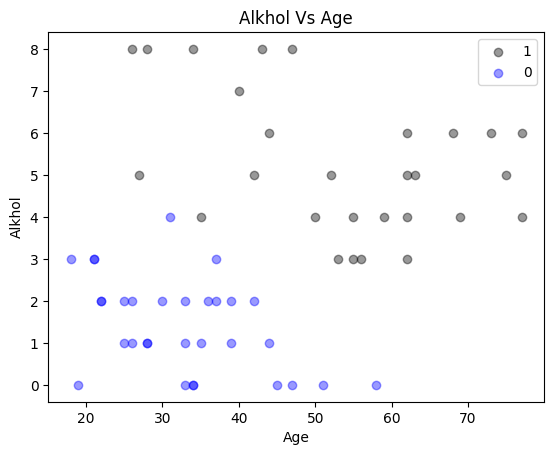
print('Classification accuracy on test', (accu\_test)\*100)

**Runtime Screenshots:**

|  |
| --- |
| **Dataset:**  **5**  **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** |







--------------Using KNN Algorithm----------------

6.855654600401044

[0 0 1 1 0 0 0 0 1 1 0 0]

Confusion Matrix:

[[7 1]

[1 3]]

In Confusion Matrix:-----

Position 1.1 shows the patients that don't have cancer

Position 1.2 shows the number of patients that have higher risk ofcancer

Position 2.1 shows the incorrect value

Position 2.2 shows the correct number of patients that have cancer

F1 Score : 75.0

Accuracy : 83.33333333333334

----------Using Decsion Tree Algorithm----------

Classification accuracy on train 95.74468085106383

Classification accuracy on test 75.0

**7.CONCLUSION**

I have processed the dataset to differentiate the affected patient and its level of the growth of the cancer by the ML system. Here it presented an approach to find best accuracy of the cancer result to assist the radiologist and the future enhancement. Further loads ought to be directed at improving the classifying accuracy levels of result through experiments with various alternatives.

A benchmarking of the most performing architectures on available datasets using similar metrics can help in their comparative analysis. Finally, one of the current limitations is the data and their imbalanced nature. The use of new loss functions designed to tackle the problem of unbalanced classes such as focal loss, could improve the existing results, and help achieve more efficient training. With more datasets and more balanced data, we think that better result can be achieved.