LUNG CANCER DETECTION

A project Report in partial fulfilment of the degree

BACHELOR OF TECHNOLOGY

In

Computer science & Engineering

By

2103A51124 Ch.Varsha

2103A51365 K.Shruthi

2103A51356 J. Akhila

Under The Guidance Of

D.Ramesh

Submitted To

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SR UNIVERSITY , ANANTHASAGAR,WARANGAL

DEPARTMENT OF COMPURTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report entitled “lung cancer Detection” is a record of bonafide work carried out by students Varsha, Shruthi Akhila bearing roll no(s) 2103A51124,2103A51365,2103A51356 during the academic year 2022-2023 in partial fulfilment of the award of the degree of **Bachelor of Technology in computer science Engineering.**

Head of the Department

MR. D. RAMESH

Asst.Professor

SR University

Ananthasagar , Warangal

ACKNOWLEDGEMENT

We express our thanks to course co-coordinator Mr. D. Ramesh,Asst .prof.

For guiding us from the beginning through the end of the course project We express our gratitude to Head of the Department CA&AI, M. sheshikala ,Associate Professor for encouragement, support and insightful suggestions. We truly value their consistence feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, school of computer science and Artificial Intelligence. Dr C.V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all teaching and non-teaching staff of the department for their suggestion and timely support.

ABSTRACT

Lung cancer is one of the most common types of cancer and is responsible for a significant number of cancer related deaths worldwide. Early detection and accurate prediction of lung cancer can significantly improve patient outcomes by allowing for earlier intervention and treatment. AI has shown great potential in the field of medical imagine and has been applied to lung cancer prediction using various imagine modalities such as chest X-rays, computed tomography (CT) scans, and magnetic resonance imaging (MRI) scans. Machine learning algorithms can analyze large volumes of medical imaging data and detect subtle patterns and abnormalities that may be missed by human observers. Deep learning, a subset of machine learning that utilizes artificial neural networks, has been particularly successful in lung cancer prediction. CNN are type of deep learning algorithm that has been applied to medical imaging and has shown promising results in in lung cancer prediction. CNN are designed to recognize patterns in image data and can be trained to identify features that are associated with lung cancer.

TABLE OF CONTENTS

1.INTRODUCTION

Lung cancer is serious and often fatal disease that affects millions of people around the world. Early detection is crucial in improving the chances of survival, but it can be challenging to detect lung cancer at early stage using traditional methods. This is where artificial intelligence (AI) comes in as a valuable tool in improving the accuracy and efficiency of lung cancer detection. AI algorithms can be trained to analyze medical images such as CT scans and X-rays, to identify potential signs of lung cancer. These algorithms are capable of detecting even the smallest anomalies that may not visible to the human eye, and they can do so quickly and accurately. One of the most promising applications of AI in lung cancer detection is the development of computer-aided diagnosis(CAD) systems .These systems use machine learning algorithms to analyze medical images and provide doctors with detailed reports highlighting potential areas of concern. This can help doctors to make more accurate diagnoses and to develop more accurate diagnoses and to develop more effective treatment plans for their patients. Another area where AI is being used in lung cancer detection is in the development of predictive models. By analyzing large datasets of patients information, including medical histories and genetic profiles, AI algorithms can identify patterns and risk factors associated with the development of lung cancer. This can help the doctors to identify high-risk patients and to implement preventative measures to reduce their risk of developing the disease.

1.1 PROBLEM STATEMENT

Lung cancer is a serious health condition that can be life-threatening if not detected and treated early.The early detection of lung cancer is crucial. For improving patients outcomes and increasing survival rates. Early detection of cancer is very important.

One way to improve early detection of lung cancer is through the use of medical imaging technology and chest X-rays, CT scans.

1.2 EXISTING SYSTEM

There are some existing system that can predict lung cancer .some of the existing systems are Early CDC lung, PulmGuard ,Paige AI which takes more time for results however this methods are not used for more accurate detection of cancer rate. Therefore there is a need for better technologies which can give accurate results for patients to save their lives.

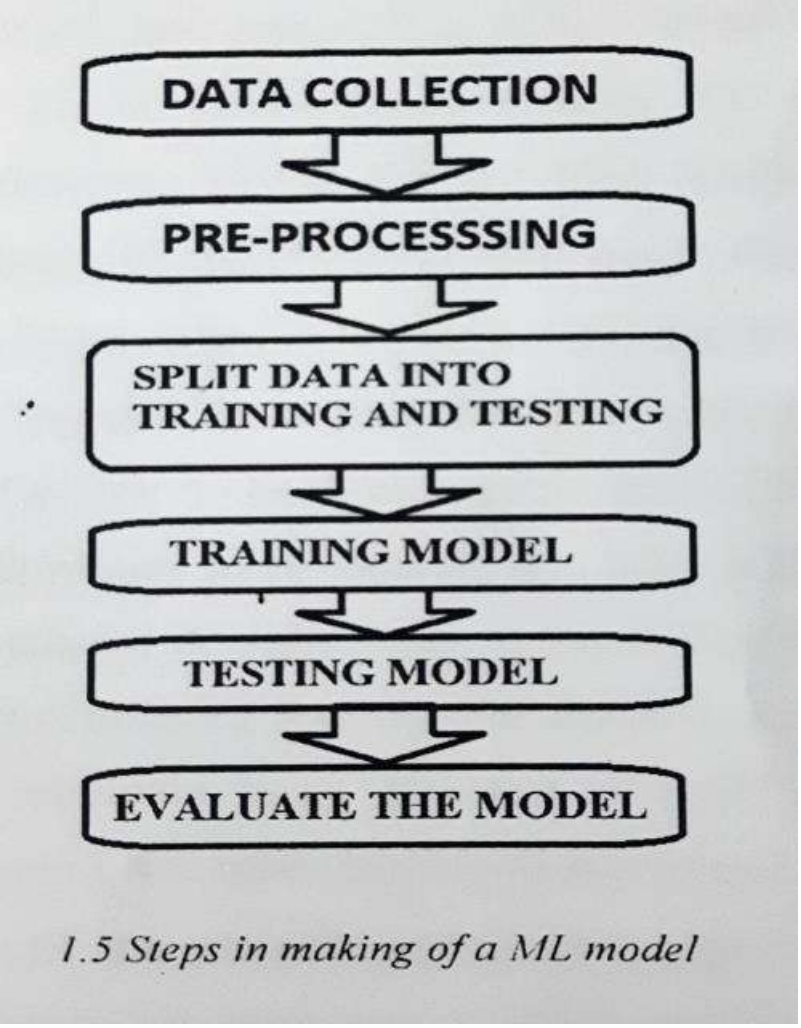
1.3 PROPOSED SYSTEM

With the assist dataset we created different algorithms , specifically logistic regression, KNN ,Decision Tree, and examine the outcomes and accuracy and find which models performs better and is reliable.

1.4 OBJECTIVES

- Compare the accuracy in 5 specific classification -based system learning algorithms
- To establish machine learning algorithms are reliable for automatic results.
- These algorithms can be used to make predictions on new data, allowing for real-time monitoring and early detection of new issues.

1.5 ARCHITECTURE



1.5 Steps in making of a ML model

CHAPTER 2

LITERATURE SURVEY

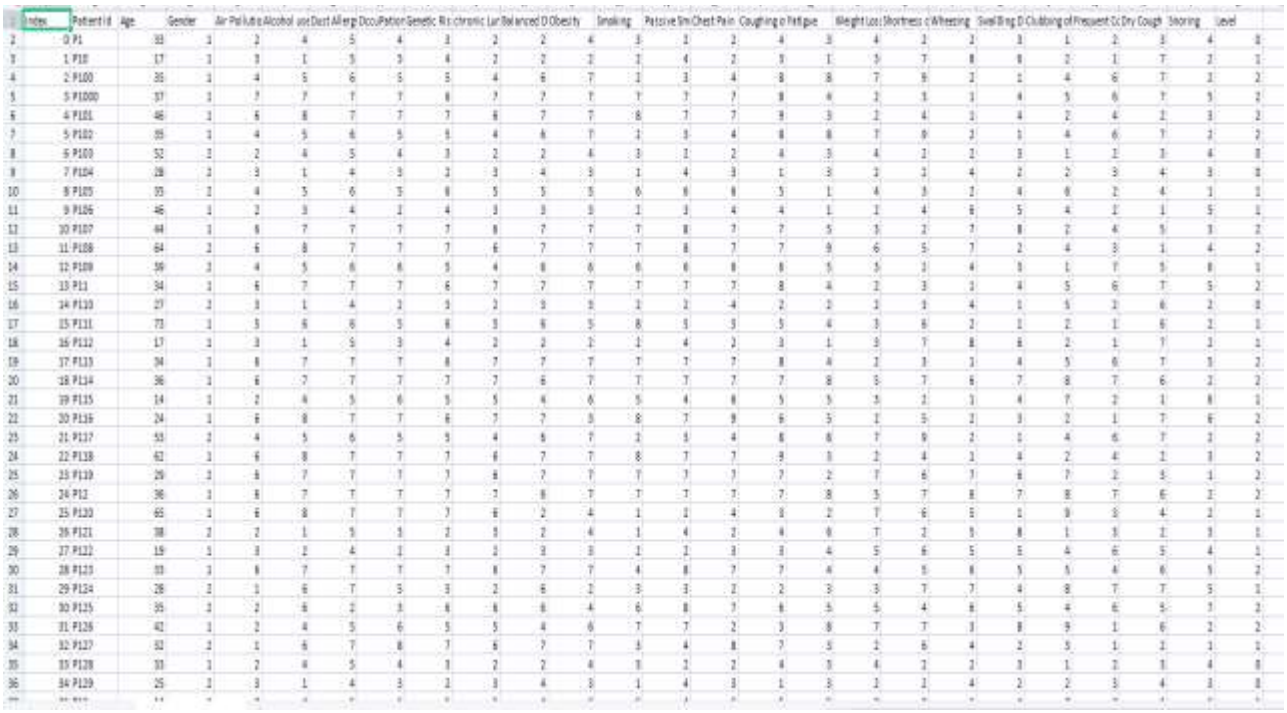
SI NO	DATE OF PUBLICATION	AUTHORS	NAME	METHODOLOGY	ACCURACY
1	1 April,2021	Marjolein a Heuvelmans	Lung cancer prediction by deep learning	SVM	94.5
2	8 August,2018	Jason L Causey,junyu Zhang	Highly accurate model for prediction of lung cancer	CNN	90.0
3	9 september 2019	Chao Zhang ,Xing Sun	Toward an expert level of lung cancer detection	CNN	84.4
4	2021	Yu Gu, Jingquian Chi	A survey of computer aided diagnosis of lung cancer.	CAD	87.5
5	2019	Yuya Onishi , Atushi Teramoto	Automated Pulmonary nodule classification	DCNN ,GAN	93.9

CHAPTER 3

DATA PREPROCESSING

Data pre-processing is an essential step in the development of Artificial Intelligence (AI) models. It involves transforming raw data into a format that is usable for machine learning algorithms. Data pre-processing is crucial because the quality of the data used in AI models determines the accuracy and reliability of the models predictions. Data pre-processing can be broken down into several stages, including data cleaning, data transformation, feature engineering and data augmentation. Each stage of the pre-processing process aims to improve the quality and accuracy of the data used in AI models. Data cleaning is the first step in the data process. It involves removing of any data that is incomplete, inconsistent or inaccurate. Data transformation is the second method that involves transforming the data into a format that is suitable for machine learning algorithms. By these pre-processing models AI provides meaningful insights from data and make accurate predictions.

3.1 DATA SET DESCRIPTION



Index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
0	P1	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	2
1	P10	17	1	3	1	5	3	4	2	2	2	3	4	2	3	1	3	7	8	6	2	1	7	2	1
2	P100	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	2
3	P1000	37	1	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	2
4	P101	46	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	2
5	P102	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	2
6	P103	52	2	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	1
7	P104	28	2	3	1	4	3	2	3	4	3	2	4	3	2	3	3	2	1	4	2	2	3	4	1
8	P105	35	1	4	5	6	5	8	5	5	5	6	8	8	5	1	4	3	2	4	8	2	4	1	1
9	P106	46	1	2	3	4	2	4	3	3	3	2	3	4	4	1	2	4	8	5	4	2	1	5	1
10	P107	44	1	6	7	7	7	7	8	7	7	7	8	7	7	5	3	2	7	8	2	4	5	3	2
11	P108	64	2	6	8	7	7	7	6	7	7	7	8	7	7	9	6	5	7	2	4	3	1	4	2
12	P109	59	2	4	5	6	6	5	4	8	6	6	6	8	8	5	5	2	4	3	1	7	5	8	1
13	P11	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	2	4	5	6	7	5	2
14	P110	37	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	0
15	P111	73	1	5	6	6	5	8	5	6	5	8	5	5	5	4	3	6	2	1	2	1	6	2	1
16	P112	17	1	3	1	5	3	4	2	2	2	1	4	2	3	1	3	7	8	6	2	1	7	2	1
17	P113	34	1	6	7	7	7	8	7	7	7	7	7	7	8	4	2	3	2	4	5	6	7	3	2
18	P114	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6	7	8	7	6	2	2
19	P115	14	1	2	4	5	6	5	5	4	6	5	4	8	5	5	5	2	3	4	7	2	1	8	1
20	P116	24	1	6	8	7	7	6	7	7	5	8	7	9	6	5	2	5	2	3	2	1	7	6	2
21	P117	53	2	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	2
22	P118	62	1	6	8	7	7	7	8	7	7	8	7	7	9	3	2	4	2	4	2	4	2	3	2
23	P119	29	2	6	7	7	7	7	8	7	7	7	7	7	7	2	7	6	7	8	7	2	5	1	2
24	P12	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	8	7	8	7	6	2	2
25	P120	65	1	6	8	7	7	7	8	2	4	1	2	4	3	2	7	6	5	1	9	3	4	2	1
26	P121	38	2	2	1	5	3	2	3	2	4	1	4	2	4	8	7	2	3	8	1	3	2	3	1
27	P122	19	1	3	2	4	2	3	2	3	3	2	2	3	3	4	5	6	5	5	4	6	5	4	1
28	P123	33	1	8	7	7	7	7	8	7	7	4	8	7	7	4	4	5	8	5	5	4	8	5	2
29	P124	28	2	1	6	7	5	3	2	6	2	3	3	2	2	3	3	7	7	4	8	7	7	5	1
30	P125	35	1	2	6	2	3	8	8	8	4	6	8	7	8	5	5	4	8	5	4	6	5	7	2
31	P126	42	1	2	4	5	6	5	5	4	6	7	7	2	3	8	7	7	3	8	9	1	6	2	2
32	P127	52	2	1	6	7	8	7	6	7	7	3	4	8	7	5	2	6	4	2	5	1	2	1	1
33	P128	33	1	2	4	5	4	3	2	2	4	3	1	2	4	3	4	1	2	3	1	2	3	4	0
34	P129	25	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	1	4	2	2	3	4	3	0

The above dataset is used to predict the lung cancer. We can detect the level of lung cancer whether it is high, medium or low. The data set includes the attributes like Age, Air pollution, Patient Id, Index, Genetic risk, occupational hazards, obesity, genetic risk, balanced diet smoking , passive , smoker , chest pain, coughing of blood . By using the above parameters we can detect the cancer easily. We used numerical values to predict the level of cancer. when the number is 0 the level is low and chance of getting cancer is low, when the number is 1 the level is medium and when the number is 2 the level is high and the chance if getting cancer is highly possible.

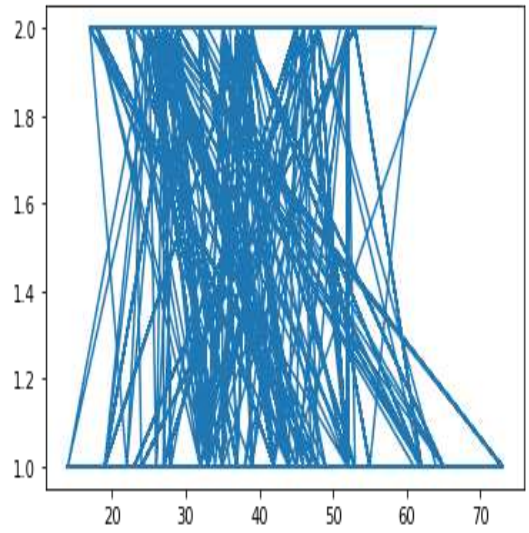
3.2 DATA PREPROCESSING THROUGH STANDARDSCALER

Data pre-processing is a crucial step in building AI model, as it involves cleaning, transformation, and preparing data in a way that makes it suitable for analysis and modelling. In our data set we used numerical 1, 2, 3 values instead of strings high, low, medium for level of prediction.

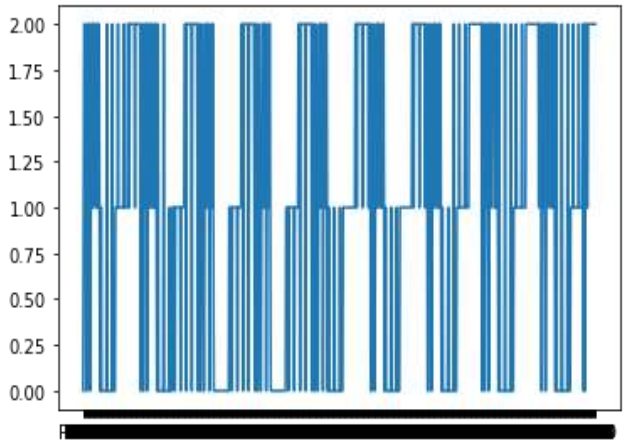
3.3 DATA VISUALIZATION

Data visualization plays a significant role in AI, as it helps in understanding complex patterns and relationships with large datasets, which may not be Immediately apparent when looking at raw data. It can help identify trends, outliers, and other patterns that are essential for effective AI modelling. In our dataset target variable is level of prediction. Our dataset includes below graphs.

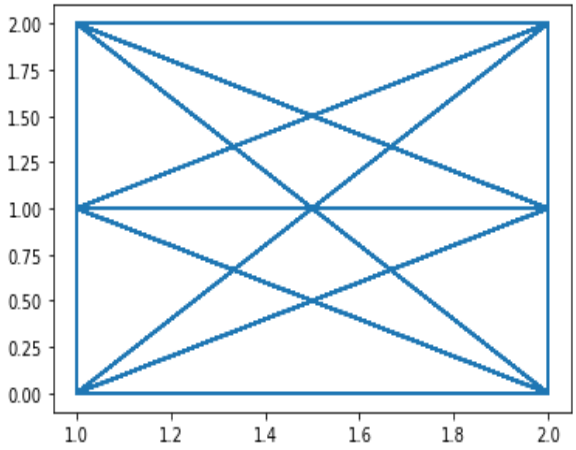
1.AGE



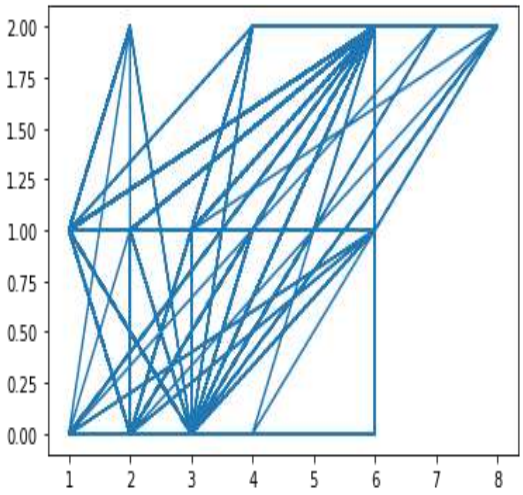
2.PATIENT ID



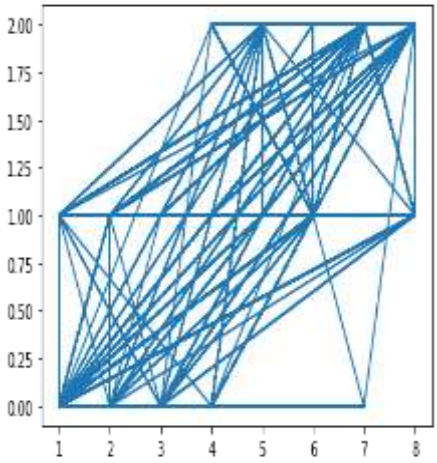
3.GENDER



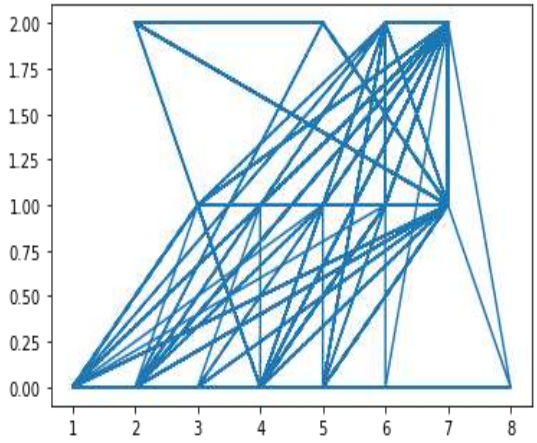
4.AIR POLLUTION



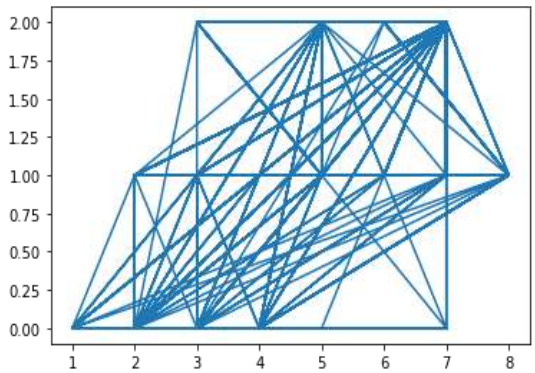
5.ALCOHOL USE



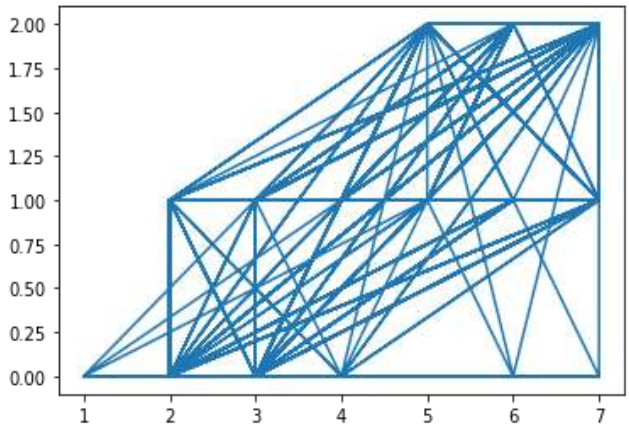
6.DUST ALLERGY



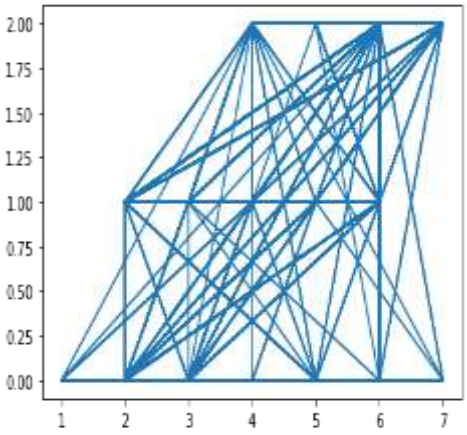
7.occupational hazards



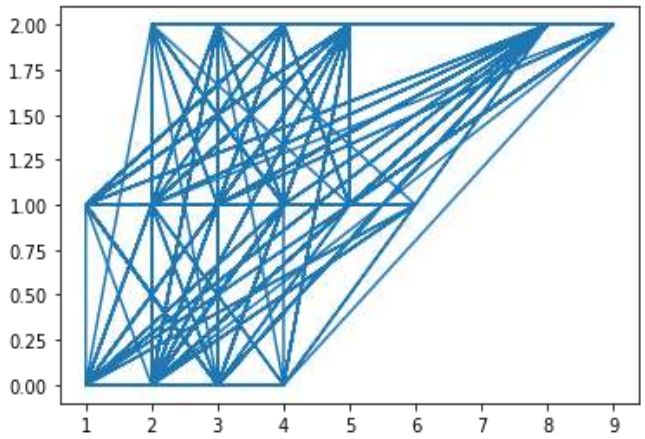
8.GENETIC RISK



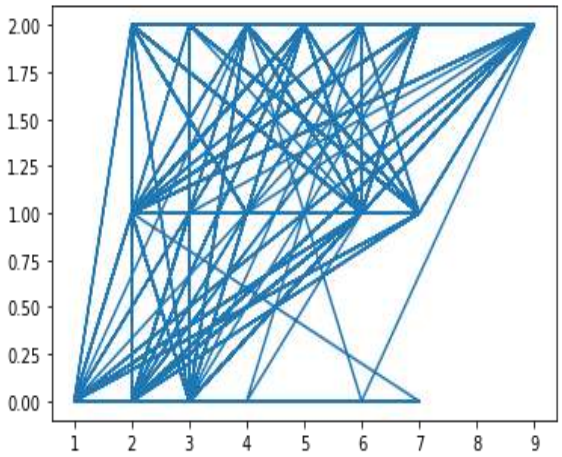
9.CHRONIC LUNG DISEASE



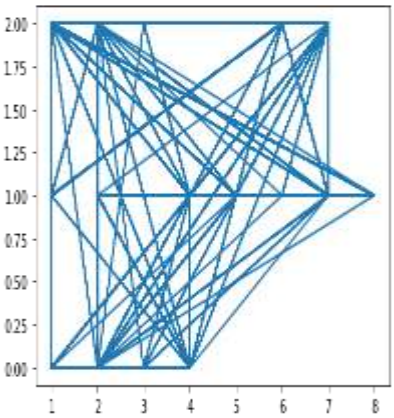
10.FATIGUE



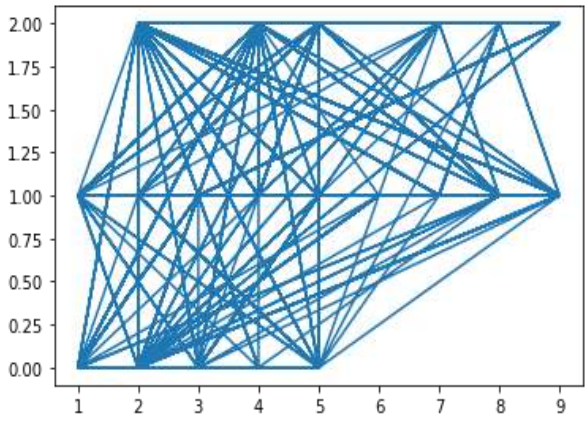
11.shortness of breathe



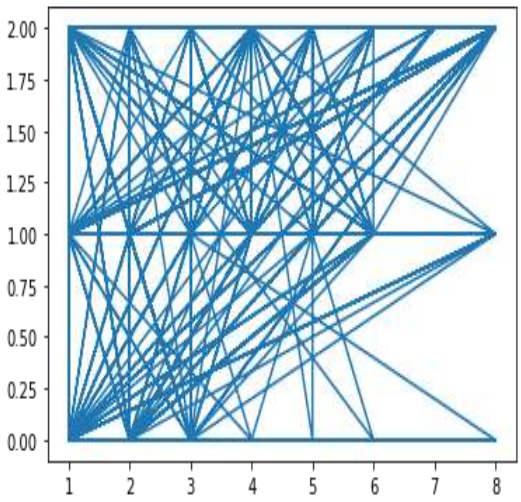
12.wheezing



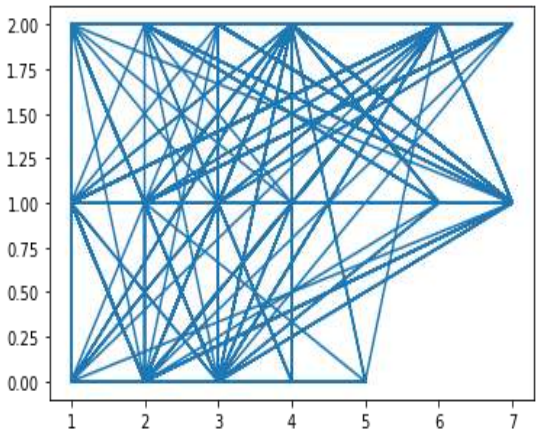
13. Swallowing difficulty



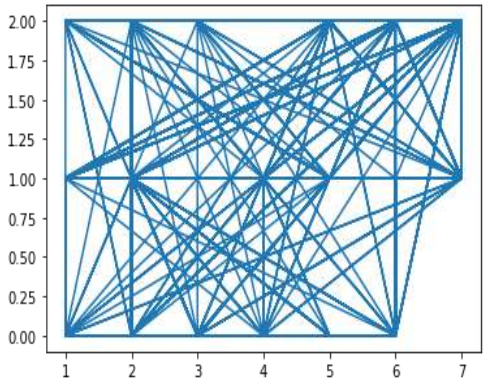
14.clubbing of fingernails



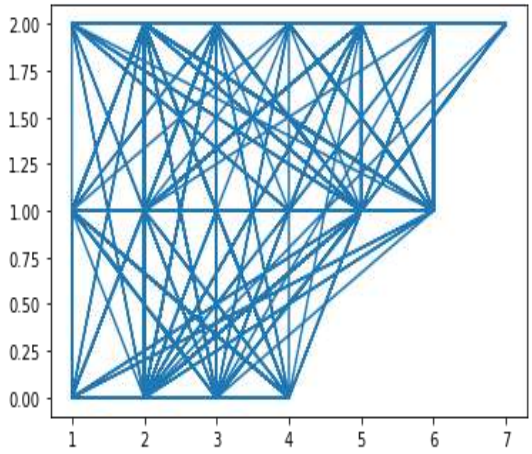
15.Frequent cold



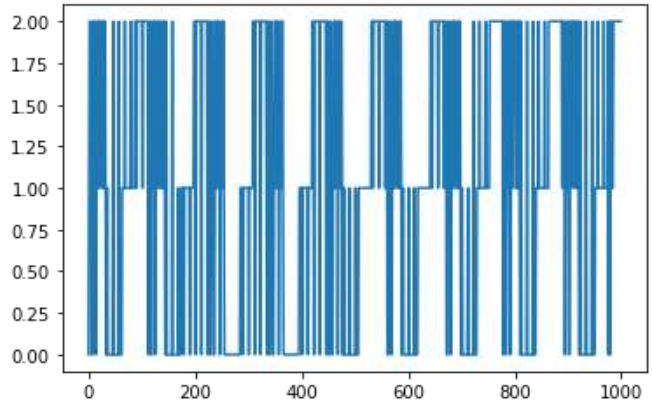
16.Dry cough



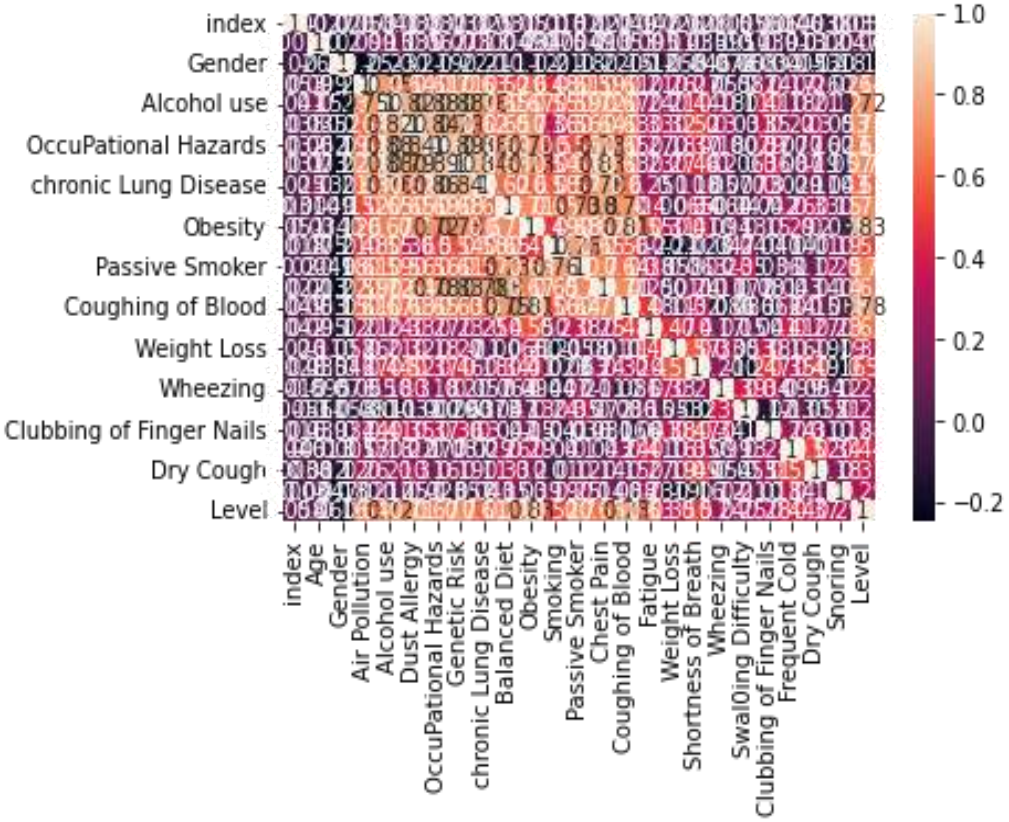
17.Snoring



18.Level

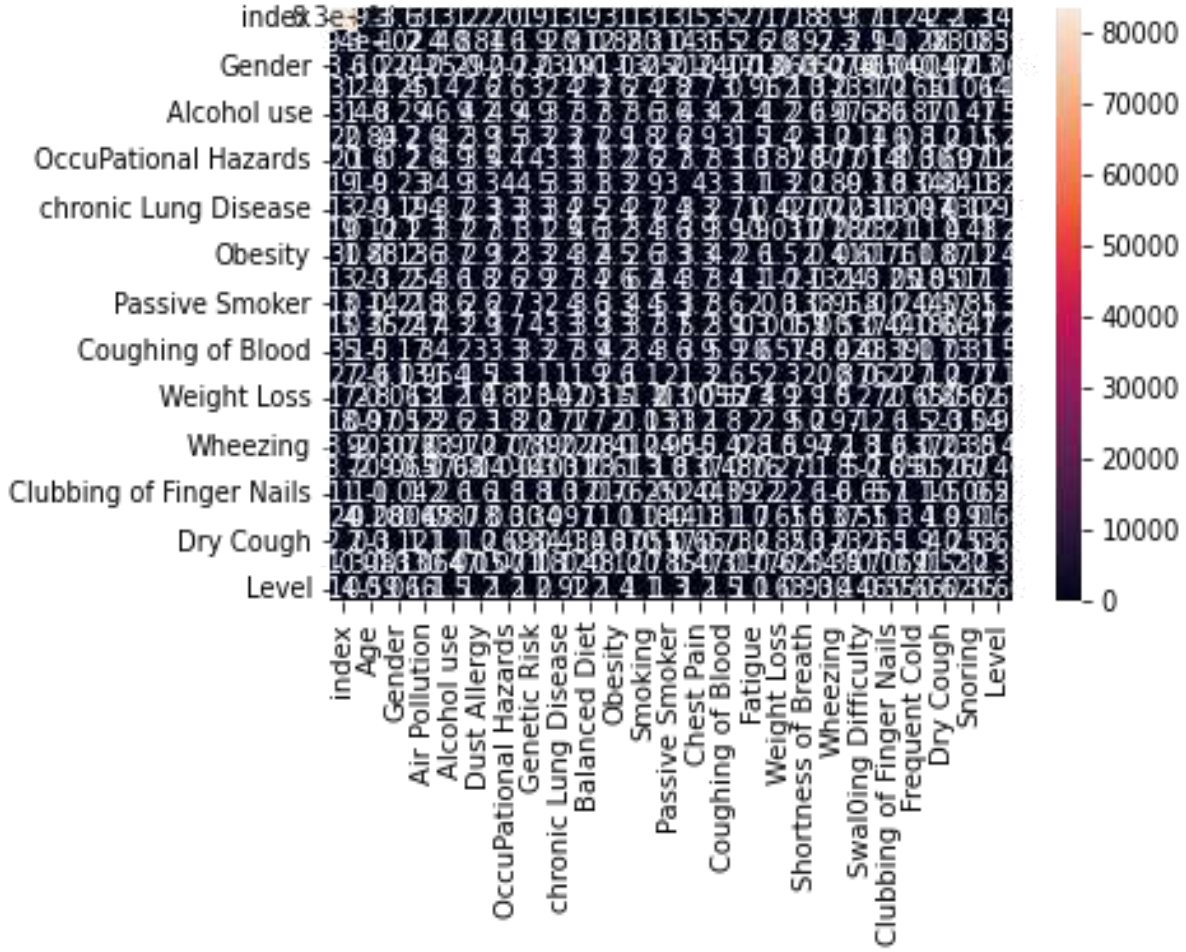


CORRELAION MATRIX



- A correlation Matrix is a table that display the correlation coefficients between multiple variable.
- A correlation coefficient 1 indicates a perfect positive correlation, coefficient 0 indicates no correlation and coefficient -1 indicates a perfect negative correlation.
- By analysing the correlation matrix , researches and analysts can identify which factors have the strongest correlation with the lung cancer and use the information to develop models or interventions to improve the prediction.

COVARIANCE MATRIX



- A covariance matrix is a square matrix that contains the covariances between pairs of variables in a dataset.
- The diagonal elements of a covariance matrix represents the variance of each variable , while the off-diagonal elements represents the covariance between the pair of variables.

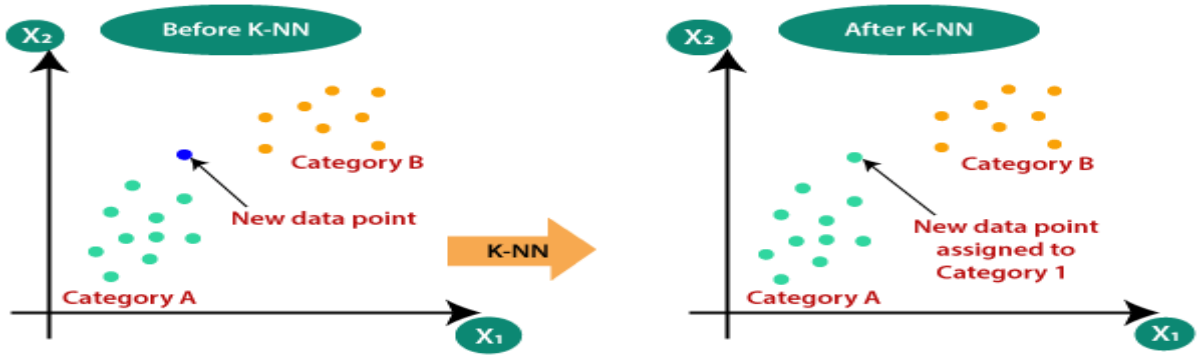
4.METHODOLOGY

4.1 KNN

KNN is a machine learning algorithm used for classification and regression tasks. In KNN, the output prediction for a given input is based on class (for classification) or values (for regression) of its K nearest neighbours in the training dataset. In KNN the distance between two points is used to determine their similarity. The value of K is typically chosen by the user and determines the number of neighbours to consider when making a prediction. To make a prediction for a new dataset point, the algorithm searches for the k nearest neighbours in the training dataset based on their distance from the new point. It then assigns the new point the class or value that is most common among its K nearest neighbours or the average value of its K nearest neighbour KNN is a simple and easy-to-implement algorithm, but it can be computationally expensive for large datasets and high-dimensional feature spaces. It is also sensitive to the choice of distance metric and the value of K.

CODE

```
from sklearn . neighbors import KNeighborsClassifier
classifier=KNeighborsClassifier(n_neighbors=5,metric='minkowski',p=2)
classifier.fit(x_train,y_train)
```



4.2 LOGISTIC REGRESSION

Logistic Regression is a statistical method used for binary classification,where the response variable has only two possible outcomes, usually represented as 0 or 1.The method is used to model the probability of a binary response variable based on one or more predictor variables.In logistic regression , the response variable is modeled as a function of the predicted variables using a logistic or sigmoid function. The logistic function takes any real-valued input and outputs a variable between 0 and 1, which can be interpreted as a probability. The logistic regression model estimates the coefficients of the predictor variables, which can be used to make predictions an new data.

CODE

```
from sklearn.linear_model import LogisticRegression

lr=LogisticRegression()

mm=lr.fit(x_train,y_train)
```

4.3 DECISION TREE

A decision tree is a type of supervised learning used in machine learning algorithm and artificial intelligence.it is a graphical representation of all the possible solutions to a decision based on certain conditions or attributes. The decision tree takes in a set of inpt data and then creates a tree-like model of decisions and their potential consequences based on the input. In decision tree each internal node represents a test on an attributes, each branch represents the outcome of the test, and each leaf node represents a class label or a decision . The tree is constructed recursively, and each decision is made based on the best split of the data at that node. Desion tree popula because they are easy to interpret and visualize , and they can handle both categorical and numerical data. They are also used for classification and regression asks, as well as for feature selection ans data mining.

CODE:

```
From sklearn.tree import DecisionTreeClassifier
classifier=DecisionTreeClassifier(criterion='entropy',
random_state=0)
mm=classifier.fit(x_train,y_train)
```

4.4 SVM

Support Vector Machine(SVM) is a powerful supervised learning algorithm used in machine learning and artificial intelligence for classification and regression analysis. SVM is based on the concept of finding the best boundary that separates the data points into different classes. The basic idea behind SVM is to find a hyperplane that separate the data points into two classes, such that the margin between the classes is maximized. The margin is defined as the distance between te hyperplane and the closest data points of each class. SVM seeks to finds the hyperplane that has the largest margin, which helps in generalizing of overfitting. In case where a linear boundary cannot separate the data points , SVM uses a technique called kernel trick to map the data inyto a higher-dimensional space, where it can be separated by a linear boundary . This technique enables SVM to handle non-linearly separable data.

CODE:

```
from sklearn.svm import SVC

svm_model=SVC(kernel='linear')

svm_model.fit(x_train,y_train)
```

4.5 GAUSSIAN NAÏVE BAYES(GNB)

Gaussian Naïve Bayes (GNB) is atype of Naïve Bayes algorithm that is used for classification tasks in machine learning and artificial intelligence . It is based on Bayes's theorem and the assumption of the independence between the features of the data. In Guassian Navive Bayes, the probability of a given input belonging to a particular class is calculated by combining the prior probability of the features given class .The algorithm assumes that the features are normally distributed , which means that they follow a Guassian dustribution.

Code:

```
from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(x_train,y_train)
```

CHAPTER 5

RESULTS

Logistic Regression Confustion Matrix

```
[2 1 1 4 7 2 2 3 5 5 7 2 7 2 4 2 7 1 7 4 5 3 1 4 1 3 5 7 3 7 5 1 1 7 5 2 7
 5 6 1 2 5 5 7 6 6 7 7 5 2 4 1 1 5 5 1 1 5 2 5 4 2 2 2 1 7 3 5 2 4 4 7 4 1
 4 1 3 7 3 1 2 2 1 5 2 6 1 7 5 4 4 4 2 3 7 4 1 5 5 6 1 7 1 7 6 5 7 4 3 5 1
 1 5 2 5 3 5 5 4 4 7 1 2 3 4 2 5 3 2 1 1 4 3 7 7 1 2 7 5 2 4 1 5 1 1 1 5 2
 1 2 4 7 5 5 5 2 2 1 2 1 7 1 3 7 2 1 2 2 7 1 1 1 3 5 7 7 5 1 2 3 7 1 1 1 3
 2 5 1 2 5 1 5 1 2 1 4 5 2 1 7 6 4 7 5 7 5 1 7 5 2 4 5 4 7 7 7 5 1 4 2 1 7
 1 2 3 5 7 4 2 4 7 5 1 2 7 6 1 2 5 4 1 7 2 7 1 5 4 1 5 6 1 5 5 4 7 1 2 5 1
 7 7 1 1 3 7 5 7 7 1 5 2 7 4 7 5 7 5 5 5 1 2 1 5 4 5 7 2 1 7 7 2 5 6 1 4 1
 7 2 1 7 1 7 5 5 1 1 6 5 5 3 5 7 7 7 7 7 2 1 4 5 7 3 7 4 4 7 1 4 1 2 2 2 7
 5 7 1 4 4 4 1 7 1 6 2 7 4 5 1 4 1 5 5 7 1 5 1 3 2 2 1 2 3 2 1 7 1 1 1 7 2
 1 5 2 1 1 7 7 3 1 2 4 2 7 1 1 1 7 2 4 1 4 2 4 1 5 7 7 7 1 1 2 3 1 7 5 2 2
 2 1 4 5 4 2 2 1 1 5 5 1 2 2 2 5 4 2 4 4 5 7 1 5 2 1 4 1 1 2 5 7 7 5 1 5 7
 5 7 5 1 1 3 7 2 2 6 6 1 1 4]
```

Accuracy: 0.6615720524017468

Decision Tree Confusion Matrix

```
[4 1 3 4 7 2 3 1 5 5 7 2 7 4 2 3 7 1 6 4 5 3 1 2 1 1 5 7 3 7 5 1 1 7 6 3 7
 5 6 3 3 5 5 6 6 6 7 7 5 2 2 1 1 5 5 1 3 5 2 5 2 4 4 4 1 7 1 5 2 2 4 7 2 1
 4 1 3 7 3 3 3 3 1 5 2 6 3 6 5 4 4 2 2 3 6 4 1 5 5 6 1 7 1 7 6 5 7 2 3 5 1
 1 5 2 5 3 5 5 4 4 7 1 4 3 4 2 5 3 2 1 1 2 3 7 7 1 4 7 5 4 2 1 5 1 1 3 5 2
 1 2 4 7 5 5 5 3 4 1 2 1 7 1 3 7 4 1 2 4 6 3 1 1 3 6 7 7 6 1 2 3 7 1 1 3 3
 2 5 1 2 5 1 5 3 2 1 4 6 2 1 7 3 4 7 5 7 5 3 7 5 3 2 5 4 7 7 7 5 1 4 3 1 6
 1 2 3 5 7 6 4 4 7 5 1 3 7 6 1 3 5 2 1 7 4 7 1 5 2 3 5 6 1 5 5 2 7 2 4 5 1
 7 7 1 1 3 6 5 7 7 3 5 2 7 2 7 5 7 5 5 5 3 2 1 5 4 5 7 4 1 7 7 4 5 6 1 2 1
 7 3 1 7 1 7 5 5 1 1 6 5 5 3 5 7 7 7 6 7 4 1 4 5 7 3 7 2 2 7 3 2 1 3 2 4 7
 5 7 1 4 4 4 1 7 1 3 2 6 2 5 1 2 2 5 5 7 1 5 1 3 3 2 1 4 3 6 1 7 1 3 1 7 2
 1 5 4 1 1 7 7 3 3 4 2 2 7 3 3 1 7 3 2 1 2 4 4 3 5 7 6 7 1 3 4 3 1 7 5 6 3
 6 1 3 6 2 2 4 1 1 5 5 1 3 2 3 6 4 2 4 2 5 6 1 5 2 1 4 1 1 4 5 6 7 5 1 5 7
 5 7 5 1 1 3 6 4 4 6 6 1 1 4]
```

Accuracy: 0.8864628820960698

KNN Confusion Matrix

```
[4 1 3 4 7 2 3 1 5 5 7 2 7 4 2 2 7 1 6 4 5 3 1 2 1 1 6 7 3 7 6 1 1 7 6 3 7
 5 6 3 3 5 5 6 6 6 7 7 5 2 2 1 1 5 5 1 3 5 2 5 2 4 4 4 1 7 1 6 2 2 4 7 2 1
 4 1 3 7 3 3 3 3 1 5 2 6 3 6 6 4 4 2 2 3 6 4 1 5 5 6 1 7 1 7 6 5 7 2 3 5 1
 1 5 2 5 3 5 5 4 4 7 1 4 3 4 2 5 3 2 1 1 2 3 7 7 1 4 7 5 4 2 1 5 1 1 3 5 2
 1 2 4 7 6 5 5 3 4 1 2 1 7 1 3 7 4 1 2 4 6 3 1 1 3 6 7 7 6 1 2 3 7 1 1 3 3
 2 5 1 2 5 1 6 3 2 1 4 6 2 1 7 3 4 7 5 7 5 3 7 5 3 2 5 4 7 7 7 6 1 4 3 1 6
 1 2 3 5 7 6 4 4 7 5 1 3 7 6 1 3 5 2 1 7 4 7 1 5 2 3 5 6 1 5 5 2 7 2 4 5 1
 7 7 1 1 3 6 5 7 7 3 5 2 7 2 7 6 7 5 5 6 3 2 1 5 4 5 7 4 1 7 7 4 5 6 1 2 1
 7 3 1 7 1 7 6 5 1 1 6 5 5 3 6 7 7 7 6 7 4 1 4 6 7 3 7 2 2 7 3 2 1 3 2 4 7
 5 7 1 4 4 4 1 7 1 3 2 6 2 6 1 2 2 5 5 7 1 6 1 3 3 2 1 4 3 6 1 7 1 3 1 7 2
 1 5 4 1 1 7 7 3 3 4 2 2 7 3 3 1 7 3 2 1 2 4 4 3 5 7 6 7 1 3 4 3 1 7 5 6 3
 6 1 3 6 2 2 4 1 1 6 5 1 3 2 3 6 4 2 4 2 5 6 1 5 2 1 4 1 1 4 5 6 7 5 1 5 7
 6 7 5 1 1 3 6 4 4 6 6 1 1 4]
```

Accuracy: 0.8842794759825328

SVM Confusion Matrix

```
[2 1 3 4 7 2 3 1 5 5 7 2 7 2 2 2 7 1 7 4 5 3 1 4 1 1 5 7 3 7 5 1 1 7 6 2 7
 5 6 1 2 5 5 7 6 6 7 7 5 2 4 1 1 5 5 1 3 5 2 5 4 2 2 2 1 7 1 5 2 4 4 7 2 1
 4 1 2 7 3 3 2 3 1 5 2 6 3 7 5 4 4 4 2 3 7 4 1 5 5 6 1 7 1 7 6 5 7 4 3 5 1
 1 5 2 5 3 5 5 4 4 7 1 2 3 4 2 5 3 2 1 1 4 3 7 7 1 2 7 5 2 2 1 5 1 1 3 5 2
 1 2 4 7 5 5 5 2 2 1 2 1 7 1 3 7 2 1 2 2 7 1 1 1 3 6 7 7 6 1 2 3 7 1 1 1 3
 2 5 1 2 5 1 5 3 2 1 4 6 2 1 7 2 4 7 5 7 5 1 7 5 2 2 5 4 7 7 7 5 1 4 2 1 7
 1 2 3 5 7 4 2 4 7 5 1 2 7 6 1 2 5 2 1 7 2 7 1 5 4 3 5 6 1 5 5 2 7 1 2 5 1
 7 7 1 1 3 7 5 7 7 3 5 2 7 2 7 5 7 5 5 5 3 2 1 5 4 5 7 2 1 7 7 2 5 6 1 2 1
 7 2 1 7 1 7 5 5 1 1 6 5 5 2 5 7 7 7 7 7 2 1 4 5 7 2 7 4 4 7 1 4 1 2 2 2 7
 5 7 1 4 4 4 1 7 1 2 2 7 4 5 1 4 1 5 5 7 1 5 1 3 2 2 1 2 3 2 1 7 1 3 1 7 2
 1 5 2 1 1 7 7 3 1 2 4 2 7 1 3 1 7 3 2 1 4 2 4 3 5 7 7 7 1 3 2 3 1 7 5 2 2
 2 1 4 6 4 2 2 1 1 5 5 1 2 2 3 6 4 2 4 4 5 7 1 5 2 1 4 1 1 2 5 7 7 5 1 5 7
 5 7 5 1 1 3 7 2 2 6 6 1 1 4]
```

Accuracy : 0.7248908296943232

Gaussian Naïve Bias Confusion Matrix

```
[2 1 1 4 7 2 1 1 5 5 7 2 6 2 3 2 7 1 7 4 5 1 1 4 1 1 5 7 3 7 5 1 1 7 6 2 7
 5 3 1 2 5 5 7 3 6 7 7 5 2 4 1 1 5 5 1 1 5 4 5 4 2 2 2 1 7 1 5 2 4 4 6 4 1
 4 1 1 7 3 1 2 1 1 5 2 6 1 7 5 4 4 4 2 1 7 4 1 5 5 3 1 7 1 7 3 5 7 4 3 5 1
 1 5 2 5 3 5 5 4 4 7 1 2 3 4 2 5 1 2 1 1 4 3 7 7 1 2 7 5 2 3 1 5 1 1 1 5 2
 1 2 4 7 5 5 5 2 2 1 4 1 7 1 3 7 2 1 2 2 7 1 1 1 1 6 7 7 6 1 2 6 7 1 1 1 3
 2 5 1 2 5 1 5 1 2 1 4 6 2 1 7 3 4 6 5 7 5 1 7 5 2 4 5 4 7 7 7 5 1 4 2 1 7
 1 2 3 5 7 3 2 4 7 5 1 2 7 6 1 2 5 4 1 7 2 7 1 5 4 1 5 3 1 5 5 4 7 1 2 5 1
 7 7 1 1 1 7 5 7 7 1 5 2 7 2 7 5 6 5 5 5 1 2 1 5 4 5 7 2 1 7 6 2 5 3 1 3 1
 7 2 1 7 1 7 5 5 1 1 3 5 5 1 5 7 7 7 7 7 2 1 4 5 7 1 7 4 4 7 1 4 1 2 2 2 7
 5 7 1 4 4 4 1 7 1 3 2 7 4 5 1 4 1 5 5 7 1 5 1 6 2 2 1 2 3 2 1 6 1 1 1 7 4
 1 5 2 1 1 6 7 3 1 2 4 2 6 1 1 1 6 1 3 1 4 2 4 1 5 7 7 7 1 1 2 6 1 7 5 2 2
 2 1 4 6 4 2 2 1 1 5 5 1 2 2 1 6 4 2 4 4 5 7 1 5 2 1 4 1 1 2 5 7 7 5 1 5 7
 5 7 5 1 1 3 7 2 2 6 3 1 1 4]
```

Accuracy: 0.6135371179039302

CHAPTER 6

CONCLUSION AND FURURE SCOPE

6.1 conclusion

Finally, after performing all the steps needed to get the results from preparation to pre-processing to performing the models (Logistic Regression, Decision Tree, k-Nearest Neighbour Gaussian Naïve Bayes, Support Vector Machine) if we observe the percentage of accuracy of Logistic Regression Decision Tree, K-Nearest Neighbour, Gaussian Naïve Bayes, Support Vector Machine are 66.15720524017468, 88.64628820960698, 88.42794759825328,61.35371179039302,72.4890829694323.Out of Which Decision Tree model with 88.646288209698 percent accuracy performs relatively better than all other models and secondly K-Nearest Neighbour model performs better accuracy with 88.42794759825328 percentage.

REFERENCES

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=lung+cancer+prediction+using+AI&oq=#d=gs_qabs&t=1682244365685&u=%23p%3DNBO1TOjjLKgJ

https://scholar.google.com/scholar?q=related:NBO1TOjjLKgJ:scholar.google.com/&hl=en&as_sdt=0,5#d=gs_qabs&t=1682244637620&u=%23p%3DE32uHYseWYAJ

https://scholar.google.com/scholar?q=related:CHF63PywS4oJ:scholar.google.com/&hl=en&as_sdt=0,5#d=gs_qabs&t=1682245067110&u=%23p%3DxEu-6AKLvn0J

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=lung+cancer+prediction+using+AI&oq=#d=gs_qabs&t=1682244365685&u=%23p%3DNBO1TOjjLKgJ