LUNG CANCER DETECTOR

Sloveni Nayak

Date: 08/06/2023

1. Abstract:

One of the leading causes of cancer-related deaths worldwide is the Lung cancer. In this prototype report, I focused on developing a machine learning-based approach for the detection of lung cancer using the medical background of the patients. Designing an accurate and efficient system that can aid doctors in identifying potential risks of lung cancer is the goal of my product. Early detection plays a crucial role in improving patient outcomes and survival rates. We can are trying to predict whether the patient is suffering from lung cancer or not, by analyzing the various symptoms of lung cancer. Here, we can use several ML classification models like KNN, SVM, and Decision Tree.

2. Problem Statement:

Lung cancer is characterized by the uncontrolled growth of abnormal cells in the lung tissues, leading to the formation of tumors. SCLC makes up 10 to 15 percent of lung cancers and is almost always caused by smoking. It is a malignant tumor that originates in the cells of the lungs. Lung cancer can be broadly categorized into two main types:

- Non-small cell lung cancer (NSCLC)
- Small cell lung cancer (SCLC)

Chest pain, shortness of breath, wheezing, coughing up blood, tiredness, weight loss, and various changes in internal fluids can only be surfaced after proper diagnosis. How is it possible to detect this life-threatening disease with proper precaution? Here, ML can play a major role in using its various ML algorithms for detection and predict accurately.

3. Market/Customer/Business Need Assessment:

Early detection and improved detection are in the growing demand for accurate and efficient lung cancer detection models to enable early diagnosis and improve the survival rate; reducing the diagnosis errors and being cost-effective solutions to the market for such helpful models or devices is a big boon in the healthcare industry. Besides, lung cancer detection models can support ongoing research efforts aimed at better understanding the disease, identifying the risk factors, discovering new biomarkers, and evaluating novel treatment options. And when we think of insurance providers, by assessing the patient risk profile regarding lung cancer, they can determine appropriate coverage, and manage healthcare costs.

The development and implementation of robust and accurate lung cancer detection models that integrate with existing healthcare systems are scalable, provide actionable insights, and adhere to regulatory and ethical standards, is required for meeting the market, customer, or business.

4. Revised Needs Statement and Target Specifications:

- Accuracy level for the test, taking into account the type of lung cancer and the stage of the disease.

- Tests that can detect the symptoms of lung cancer are chronic disease checks, blood tests, allergies, and many more.
- Sensitivity and specificity of the test in detecting lung cancer and differentiating between different types of lung cancers.
- The model's architecture should be designed to effectively handle the complexity and variability of lung cancer detection.
- Accessibility of the test, including the cost, time required, and availability of the test in different regions.
- The model should show robustness across different datasets.

Meeting these target specifications and characterizations is essential to develop a reliable and clinically applicable lung cancer machine learning model, that can assist healthcare professionals in accurate detection, classification, and segmentation of lung cancer, ultimately improving patient care and outcomes.

5. External Search:

As we all know that the dataset plays a crucial role in ML. It employs a variety of statistical, probabilistic, and optimization techniques that allows computers to detect hard-to-discern patterns from large, noisy, or complex data sets. There are not very common proper datasets for it. We used Kaggle which has a large amount of data. The dataset is designed in such a way that is easy to understand. The results were based on the data available on the internet. Percentages were calculated roughly in a relative manner.

Symptoms of lung cancer like continuous coughing, smoking, anxiety, peer pressure, chronic disease, fatigue, and many more; along with that a target variable which is the binary class containing the result that can either be positive or negative.

Dataset:

https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

6. Benchmarking Alternative Products:

A lung cancer diagnosis is completely based on the doctor's point of view. Our product can give more assurance to the doctor's decision. This machine learning model uses a lot of datasets so it can give more accurate knowledge of similar types of patients. The only source for doctors to predict or judge lung cancer is by performing a CT scan.

Some of the following perks of using our Lung cancer trained detection device result in very easy usability with better assurance. It takes very less cost as compared to the fees taken by the hospitals. Its accuracy is always high and it could be readily available as web apps so it could be beneficial for common people resulting in saving time for people. It will be readily available and accessible on various E-commerce sites, so it will be readily available to the public.

7. Applicable Regulations:

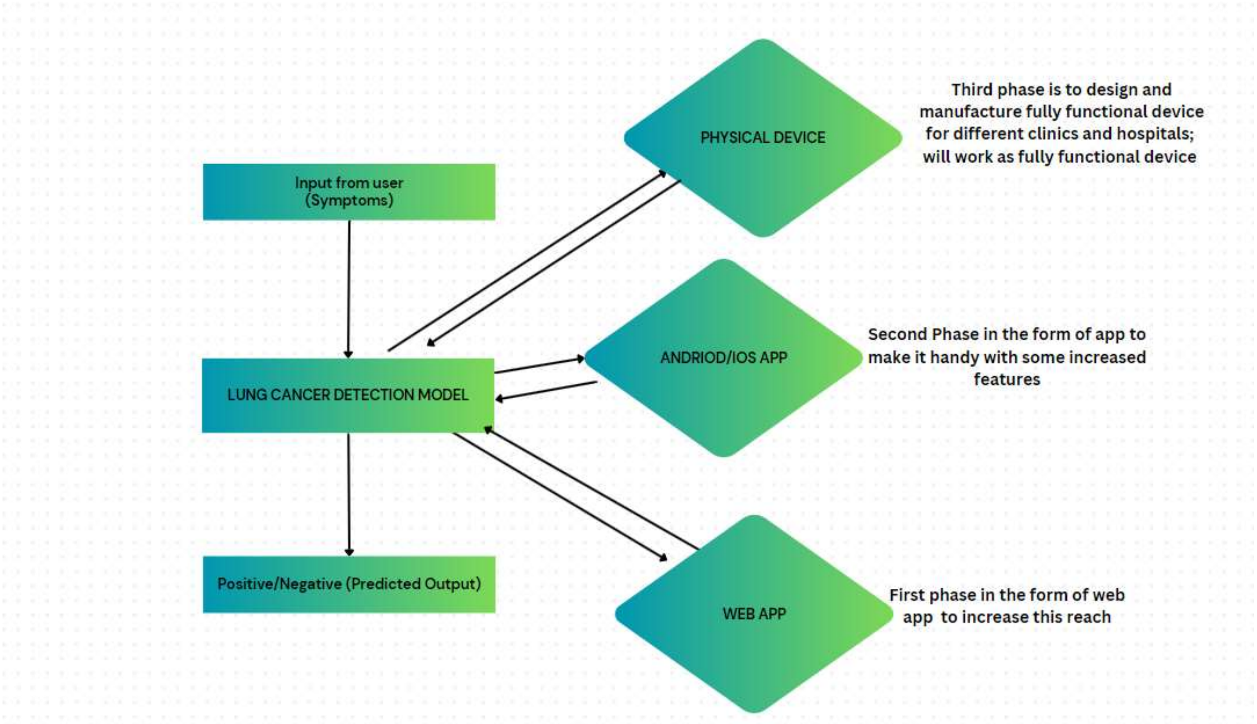
The regulations surrounding the use of machine learning in lung cancer detection devices may vary by country. In India, the regulation of medical devices is governed by the Central Drugs Standard Control Organization (CDSCO) under the Ministry of Health and Family Welfare. The following regulations are The Personal Data Protection Bill (2019) and The Indian Medical Devices Amendment Rules, (Central Drugs Standard Control Organization, 2020).

8. Applicable Constraints:

- Since the lung cancer detection device is completely based on Data, so proper availability of data is required and it must be accurate. It can be pretty challenging for the data scientists.
- ML Algorithms are complex and require extensive data and computational resources to develop and deploy effectively. Confidential health data is to be obtained to train the model.
- Devices must comply with relevant regulations like European Union's Medical Device Regulation (MDR), to ensure the safety and efficacy of the device.
- It is hard to manage the ethical regulations being a machine.
- The cost of developing, manufacturing, and deploying the device as well as the cost of training the healthcare providers to use the device is important because it can be implemented as a physical device or a web app.
- Keeping the belief that it is a machine, people might fail to show trust in this AI product until and unless it is prescribed by well-known doctors and gives perfect accuracy.

9. Business Model:

In the below-mentioned business model, I have clearly shown the 3-phase plan of the Lung Cancer Detection model. This model will support the diagnosis of various doctors along with the power of past statistics. In the first phase, we must build it over the website, since it will be starting phase so in order to monetize it, we must rely on the online advertisement. It would help us increase the reach of the product. Then we could shift this to an Android app in the second phase. Where apart from predictions we can shift this to an Android/ios app in the second phase.


PHYSICAL DEVICE
Third phase is to design and manufacture fully functional device for different clinics and hospitals; will work as fully functional device
Input from user (Symptoms)
ANDRIOD/IOS APP
Second Phase in the form of app to make it handy with some increased features
LUNG CANCER DETECTION MODEL
Positive/Negative (Predicted Output)
WEB APP
First phase in the form of web app to increase this reach

Besides the predictions, we can also suggest the most appropriate doctors with various precautionary measures. We can also add subscriptions for additional features like fixing appointments with the doctors, in the monetization part. In the end, the third phase will primarily focus on the manufacturing of devices that can be implemented in various hospitals and clinics.

On the other half, we can take a survey from the doctor's opinions which this app suggests as well as from the hospitals where it is implanted.

A good amount will be obtained through advertisement if we use a web app or Android app and we'll receive a commission of 10% to 19% if we use them in the form of a physical device.

10. Concept Generation

Machine learning not only detects lung cancer results faster but also gives higher accuracy which is around 70 to 80 % which is more significant when compared with doctors or clinicians. Lung cancer is a kind of disease which when treated early would save many lives. Therefore, ML is the most sort after technique that is very useful in replacing the present lung cancer prediction process.

A. Problem Statement Identification:

The first step to solving any problem is to analyze and identify it properly. In this case, the main issue which my device resolves is diagnosing lung cancer on the basis of various symptoms which people show like continuous coughing, yellow fingers, chronic disease, fatigue, and many more.

B. Goal & Objective:

Fixing the goal and taking respective objectives for the creation and better performance of our device. Based on symptoms and blood tests, the objectives of the device can be to give precise and accurate results.

C. Impact & Feasibility:

The required knowledge and resources along with the technological and legal hurdles must be overcome. It is possible to determine whether the proposed solution is feasible.

By evaluating the potential advantages for patients, healthcare professionals, and the healthcare system, it is possible to gauge the impact of the solution.

D. Appropriate ML Algorithms & Techniques:

Depending on the nature of the data and the goals of the device, various types of learning will be used like supervised and unsupervised learning, deep learning, NLP, or other techniques.

E. Data Requirements:

It should be defined including the type, quality, and quantity of data that is needed. This may involve the collection and labeling of new data, accessing existing datasets, or partnering with healthcare organizations to obtain the necessary data.

F. System Architecture Designation:

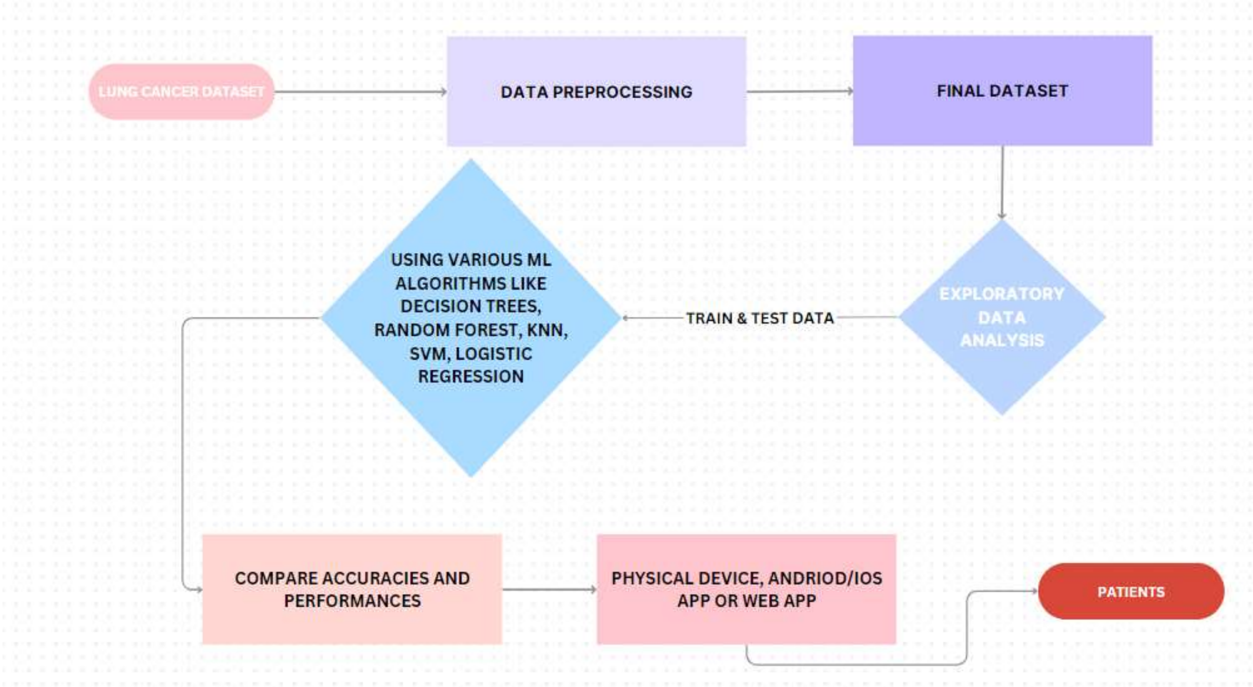
Includes a selection of hardware and software components over which our device will work. May involve cloud-based or on-premise solutions, and a selection of programming languages and frameworks. Scalability and efficiency will also matter.

G. Passing Device Prototype Through Various Tests:

Involve integrating with medical imaging and electronic health record systems, designing a UI, and the accuracy and performance validation of the device.

11. Final Product Prototype

Our Lung Cancer Detection Device is completely AI-based using an ML algorithm to analyze medical data along with patient symptoms like coughing, chronic disease, smoking, and many more. Our product tries to find out various hidden patterns or relations between various lung cancer symptoms by using various ML algorithms; also pattern recognition for human clinicians might be difficult to recognize. It will help in the healthcare system as this system will provide a quick and accurate diagnosis reducing the healthcare cost and making it accessible to public.


LUNG CANCER DATASET
DATA PREPROCESSING
FINAL DATASET
EXPLORATORY DATA ANALYSIS
TRAIN & TEST DATA
USING VARIOUS ML ALGORITHMS LIKE DECISION TREES, RANDOM FOREST, KNN, SVM, LOGISTIC REGRESSION
COMPARE ACCURACIES AND PERFORMANCES
PHYSICAL DEVICE, ANDRIOD/IOS APP OR WEB APP
PATIENTS

A multidisciplinary team of experts is required to develop such a device. It includes data analytics, data scientists, project managers, AI/ML engineers, software developers, and healthcare professionals; who select appropriate ML algorithms, label frameworks, and software, and perform data preprocessing, following the best practices for software development and deployment.

Our products are compliant, regulated & validated with valid ethical and legal standards.

What is done?

It would involve a web application to make it user-friendly and appealing. It would be able to collect medical data and symptoms and then preprocess the data. It could be designed as an on-premise or cloud-based solution according to the demand made by the organization. It would be using several classification algorithms to detect the hidden patterns and diagnose the patient more efficiently. Its scalability would be maintained to train it over a big amount of data.

On the basis of past healthcare data, a diagnosis report would give the predicted output which would also include performance metrics (accuracy, specificity, sensitivity). It would be able to maintain data privacy, security regulations, FDA approval, and other legal standards.

12. Product Details

Our product collects a dataset of various symptoms (continuous coughing, yellow fingers, smoking, anxiety, depression, peer pressure). The collected data is preprocessed to ensure its good quality and proper labeling, to be ready to use by ML algorithms. To ensure consistency, it involves removing any duplicate information, filling in missing data, and normalizing the data. Here, a proper classification algorithm is used (decision tree, KNN, logistic regression).

The model is trained over preprocessed data and then tested over the unseen data; evaluating its prediction and determining its accuracy. Finally, the model is integrated into a web application or medical device that healthcare professionals can use to assist in diagnosing lung cancer.

12.1 Data Sources

- ➔ Medical Imaging such as CT scan, ultrasound, MRI
- ➔ Blood Test Results for lung cancer detection.
- ➔ Studies made from Research papers and articles
- ➔ Survey made of patients to identify patterns
- ➔ Electronic Health Records (EHR) for lab test results and imaging data

12.2 Algorithms

Several classification algorithms like SVM (Support Vector Machine), logistic regression, decision tree, and neural networks are used to build predictive models for lung cancer prediction. Other unsupervised algorithms like K-clustering, Principal Component Analysis, and density-based clustering algorithms are used for pattern identification.

12.3 Frameworks/Libraries

Django and Flask (Python-based frameworks) are used to create Andriod App and Web Apps. ML libraries like Scikit-Learn, Tensorflow, and PyTorch are used for data preprocessing and model evaluation; DL libraries like Keras, and OpenCV for image detection and object detection.

12.4 Data Visualization

Visualization tools like Seaborn and Matplotlib are imported for data exploration and data visualization.

12.5 Software

Cloud platforms like AWS, GCP, and Microsoft Azure can be used to run heavy models. IDE used can be Google Colab, Pycharm, Jupyter Notebook, and others. Kubernatics/Docker for shipping and running applications.

12.6 Cost/Expenses

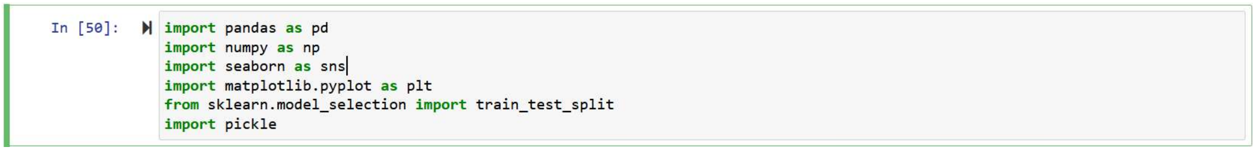
Getting large amounts of accurate medical data such as medical symptoms, and medical images can be expensive. Advanced infrastructure with GPU is required to train machine learning and deep learning model; which can cost too much for large datasets. Proper advertisement also requires a good amount of fund. Cost of having regulatory approval (FDA approval) and other legal standards for regulatory compliance can be high.

12.7 Team Required to Develop

- ➔ Healthcare professionals and medical doctors
- ➔ Machine Learning Engineers
- ➔ Cloud Engineers
- ➔ Data Scientists/Data Analysts
- ➔ Statisticians
- ➔ Project Managers
- ➔ Software Developer

13. CODE IMPLEMENTATION

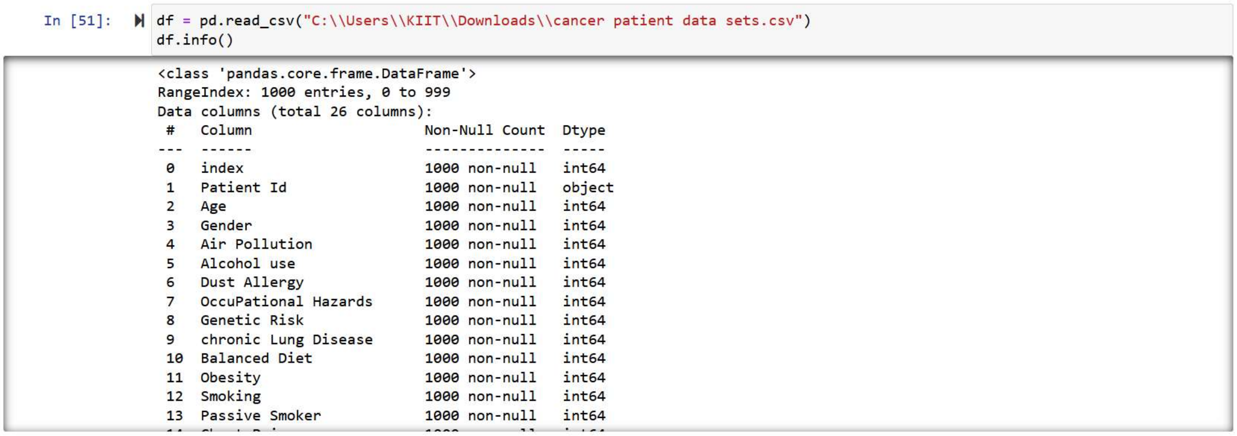
13.1. Importing Libraries



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import pickle
```

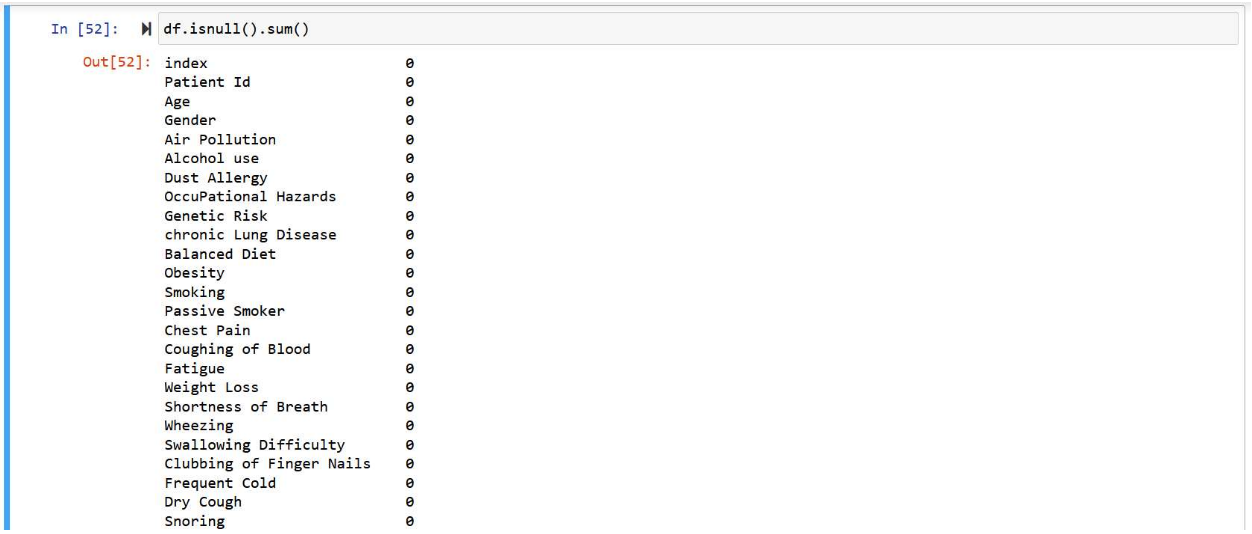
Used libraries like Pandas, NumPy, Seaborn, Scikit-Learn, Matplotlib, and many more.

13.2 Loading Dataset



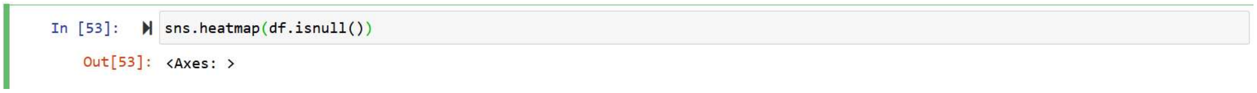
```
In [51]:  df = pd.read_csv("C:\\Users\\KIIT\\Downloads\\cancer patient data sets.csv")
          df.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 1000 entries, 0 to 999
          Data columns (total 26 columns):
           #   Column                    Non-Null Count  Dtype
          ---  ------                    --------------  -----
           0   index                     1000 non-null   int64
           1   Patient Id                1000 non-null   object
           2   Age                       1000 non-null   int64
           3   Gender                    1000 non-null   int64
           4   Air Pollution             1000 non-null   int64
           5   Alcohol use               1000 non-null   int64
           6   Dust Allergy              1000 non-null   int64
           7   OccuPational Hazards      1000 non-null   int64
           8   Genetic Risk              1000 non-null   int64
           9   chronic Lung Disease      1000 non-null   int64
           10  Balanced Diet             1000 non-null   int64
           11  Obesity                   1000 non-null   int64
           12  Smoking                   1000 non-null   int64
           13  Passive Smoker            1000 non-null   int64
```

13.3 Data Preprocessing

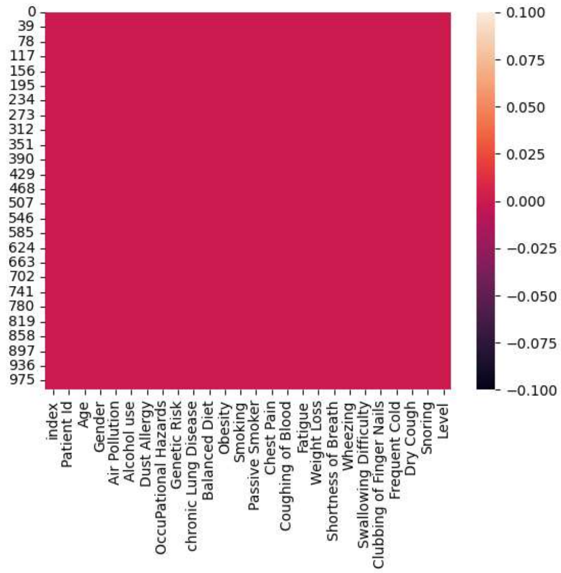


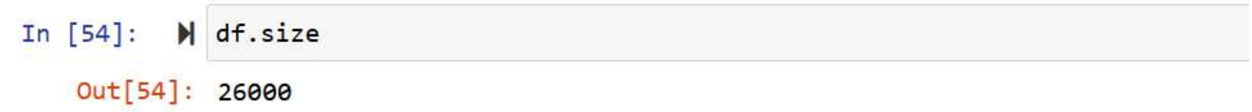
```
In [52]:  df.isnull().sum()

Out[52]:  index                       0
          Patient Id                  0
          Age                         0
          Gender                      0
          Air Pollution               0
          Alcohol use                 0
          Dust Allergy                0
          OccuPational Hazards        0
          Genetic Risk                0
          chronic Lung Disease        0
          Balanced Diet               0
          Obesity                     0
          Smoking                     0
          Passive Smoker              0
          Chest Pain                  0
          Coughing of Blood           0
          Fatigue                     0
          Weight Loss                 0
          Shortness of Breath         0
          Wheezing                    0
          Swallowing Difficulty       0
          Clubbing of Finger Nails    0
          Frequent Cold               0
          Dry Cough                   0
          Snoring                     0
```

13.4 Create Seaborn Heatmap

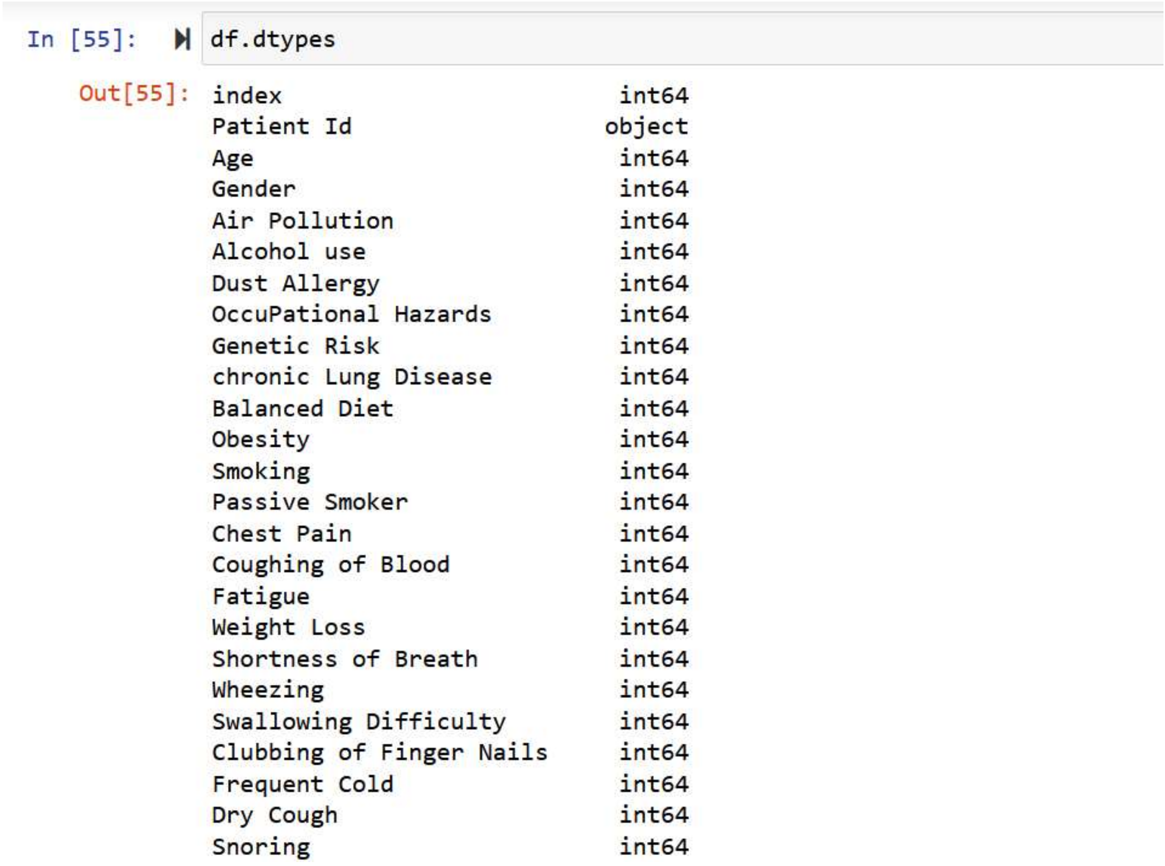


```
In [53]:  sns.heatmap(df.isnull())

Out[53]:  <Axes: >
```

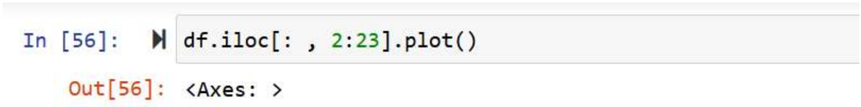




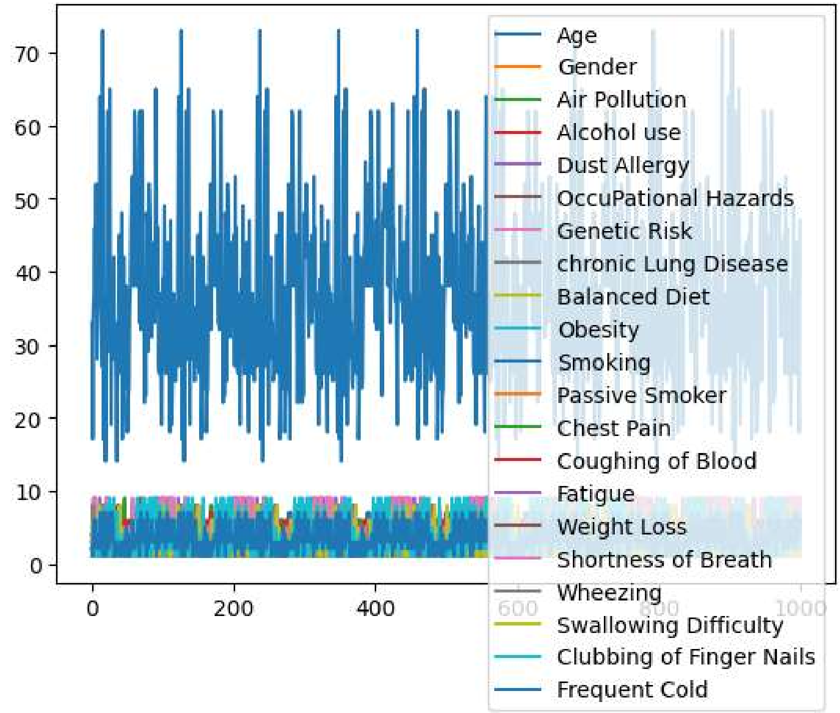
```
In [54]:  df.size

Out[54]:  26000
```



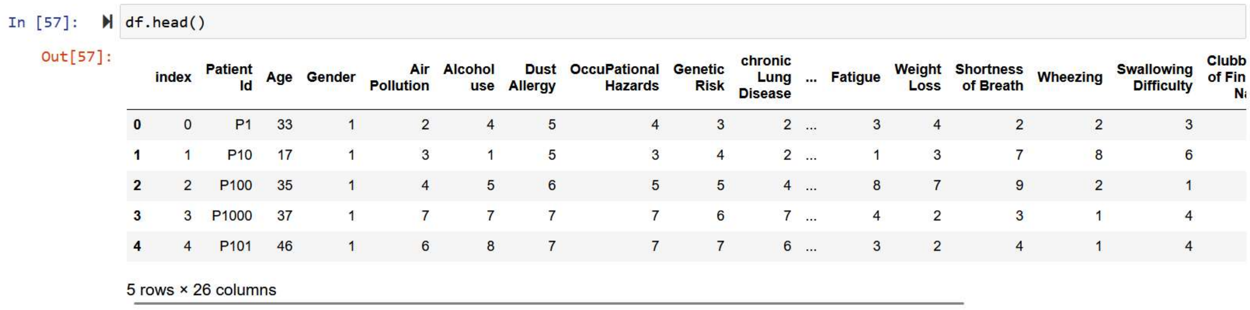
```
In [55]:  df.dtypes

Out[55]:  index                      int64
          Patient Id                object
          Age                        int64
          Gender                     int64
          Air Pollution              int64
          Alcohol use                int64
          Dust Allergy               int64
          OccuPational Hazards       int64
          Genetic Risk               int64
          chronic Lung Disease       int64
          Balanced Diet              int64
          Obesity                    int64
          Smoking                    int64
          Passive Smoker             int64
          Chest Pain                 int64
          Coughing of Blood          int64
          Fatigue                    int64
          Weight Loss                int64
          Shortness of Breath        int64
          Wheezing                   int64
          Swallowing Difficulty      int64
          Clubbing of Finger Nails   int64
          Frequent Cold              int64
          Dry Cough                  int64
          Snoring                    int64
```



```
In [56]:   df.iloc[: , 2:23].plot()

Out[56]:  <Axes: >
```



13.5 Final Dataset:

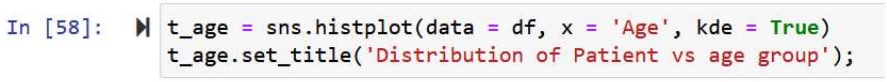


```
In [57]:   df.head()

Out[57]:
```

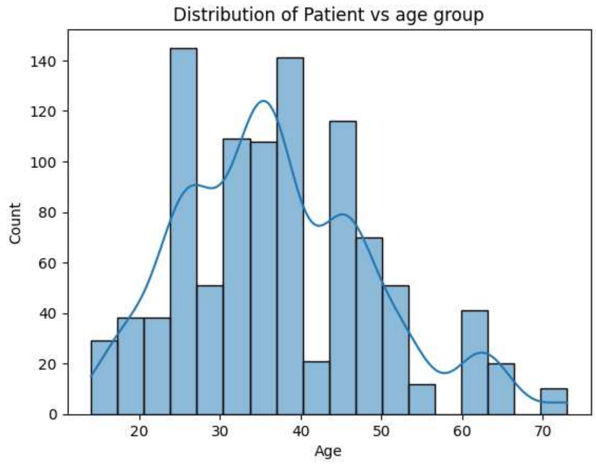
	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubb of Fin N
0	0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3	
1	1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6	
2	2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1	
3	3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4	
4	4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4	

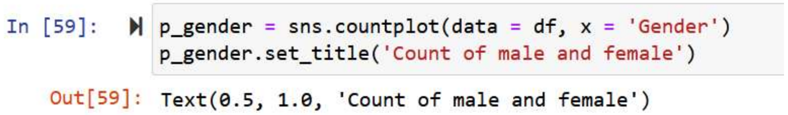
5 rows × 26 columns

13.6 Exploratory Data Analysis (EDA)

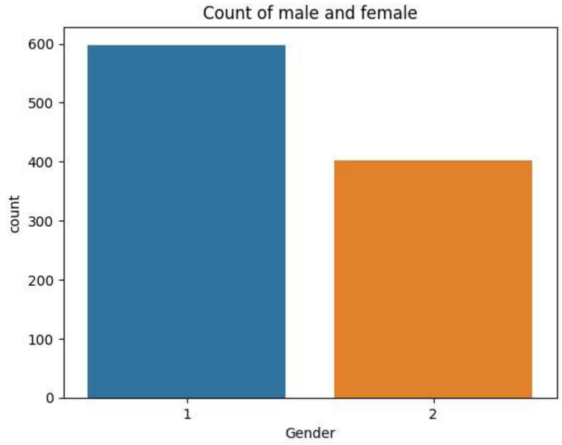


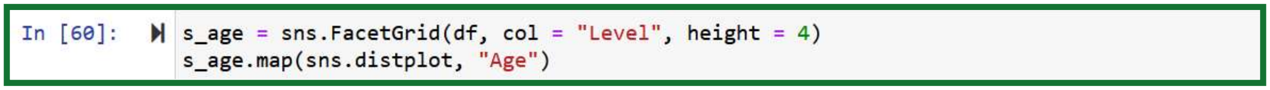
```
In [58]:  t_age = sns.histplot(data = df, x = 'Age', kde = True)
          t_age.set_title('Distribution of Patient vs age group');
```



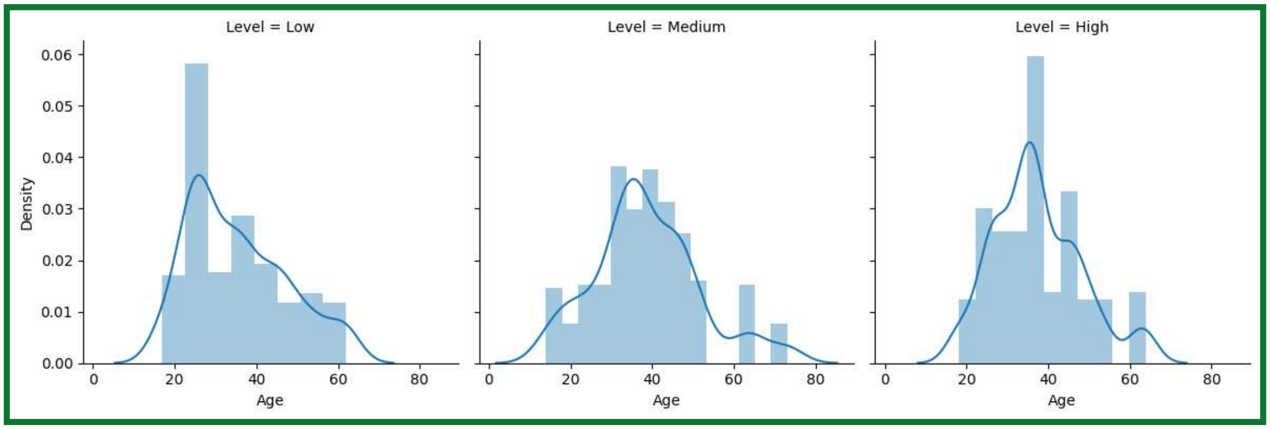


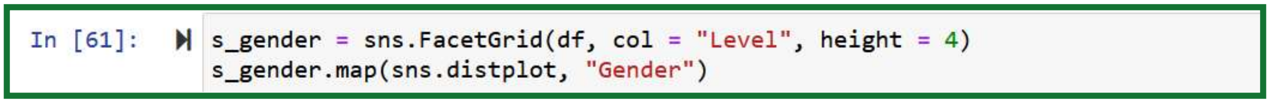
```
In [59]:  p_gender = sns.countplot(data = df, x = 'Gender')
          p_gender.set_title('Count of male and female')

Out[59]:  Text(0.5, 1.0, 'Count of male and female')
```



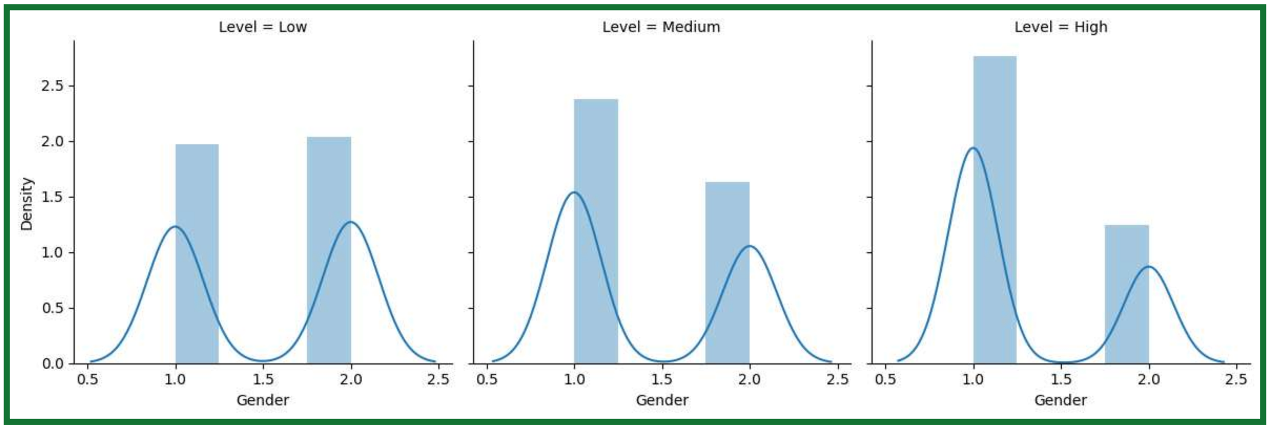


```
In [60]:  s_age = sns.FacetGrid(df, col = "Level", height = 4)
          s_age.map(sns.distplot, "Age")
```



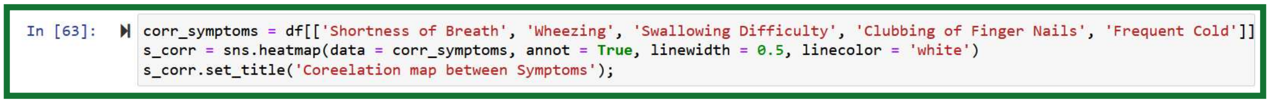


```
In [61]:  s_gender = sns.FacetGrid(df, col = "Level", height = 4)
          s_gender.map(sns.distplot, "Gender")
```

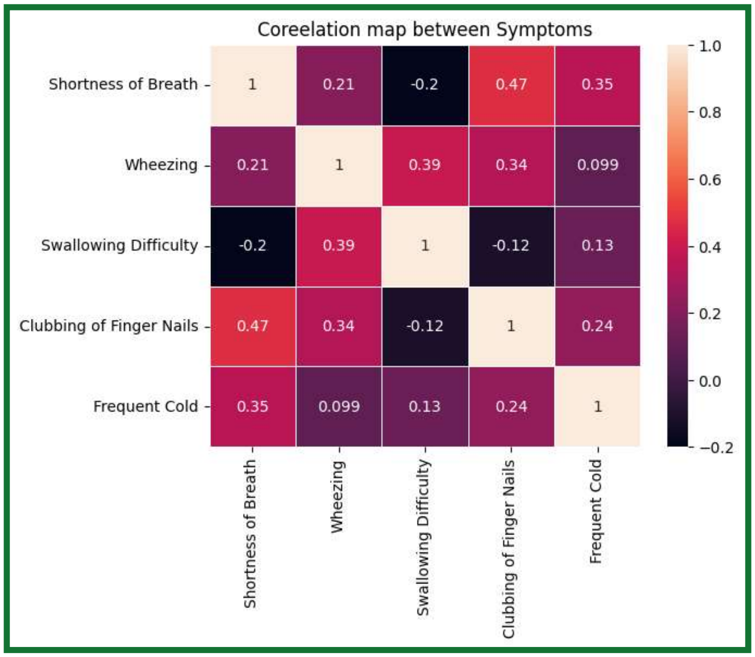


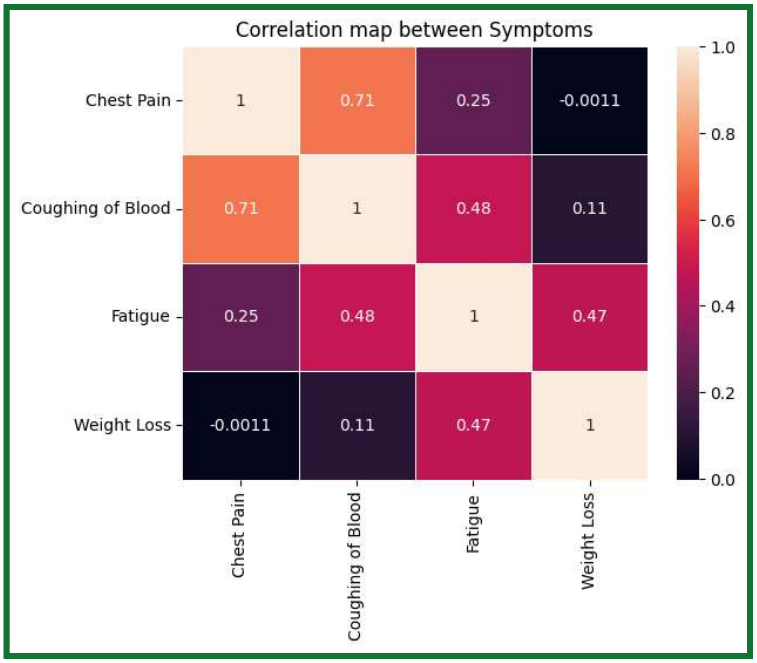
13.7 Correlation Matrix

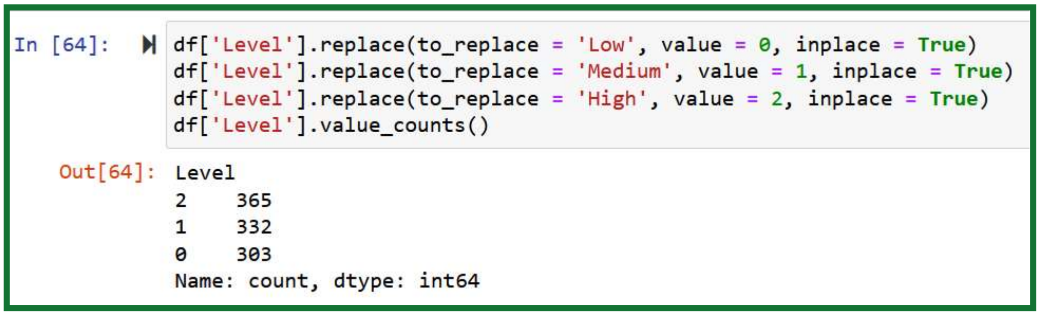
Correlation Matrix with Some Reduced Features



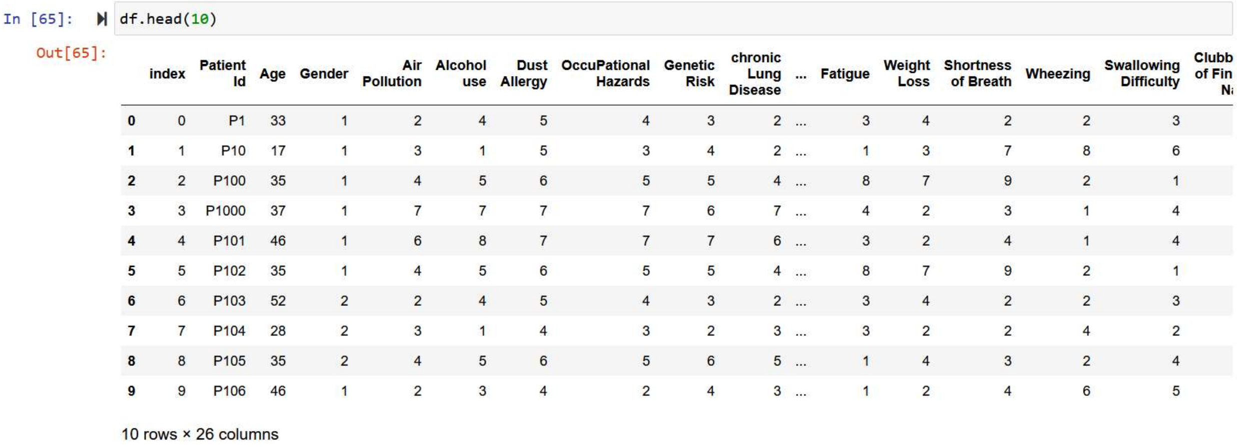
```
In [63]:  corr_symptoms = df[['Shortness of Breath', 'Wheezing', 'Swallowing Difficulty', 'Clubbing of Finger Nails', 'Frequent Cold']]
          s_corr = sns.heatmap(data = corr_symptoms, annot = True, linewidth = 0.5, linecolor = 'white')
          s_corr.set_title('Coreelation map between Symptoms');
```







```
In [64]:  df['Level'].replace(to_replace = 'Low', value = 0, inplace = True)
          df['Level'].replace(to_replace = 'Medium', value = 1, inplace = True)
          df['Level'].replace(to_replace = 'High', value = 2, inplace = True)
          df['Level'].value_counts()

Out[64]:  Level
          2    365
          1    332
          0    303
          Name: count, dtype: int64
```



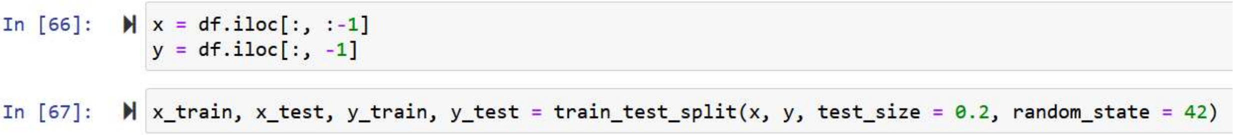
```
In [65]: df.head(10)
```

Out[65]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubb of Fin Na
0	0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3	
1	1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6	
2	2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1	
3	3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4	
4	4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4	
5	5	P102	35	1	4	5	6	5	5	4	...	8	7	9	2	1	
6	6	P103	52	2	2	4	5	4	3	2	...	3	4	2	2	3	
7	7	P104	28	2	3	1	4	3	2	3	...	3	2	2	4	2	
8	8	P105	35	2	4	5	6	5	6	5	...	1	4	3	2	4	
9	9	P106	46	1	2	3	4	2	4	3	...	1	2	4	6	5	

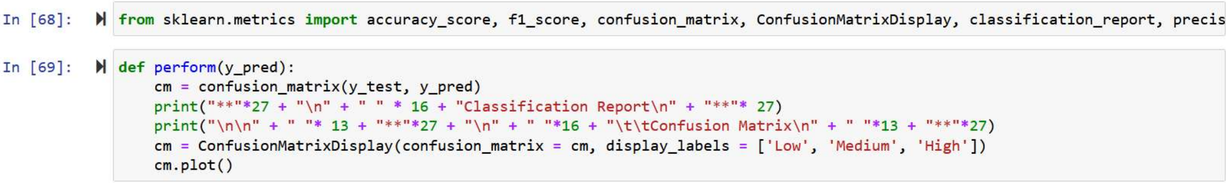
10 rows × 26 columns

13.8 Splitting The Dataset



```
In [66]: x = df.iloc[:, :-1]
         y = df.iloc[:, -1]

In [67]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

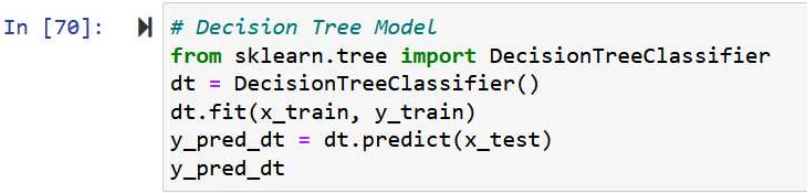
13.9 Function to check the performance and accuracy of the model



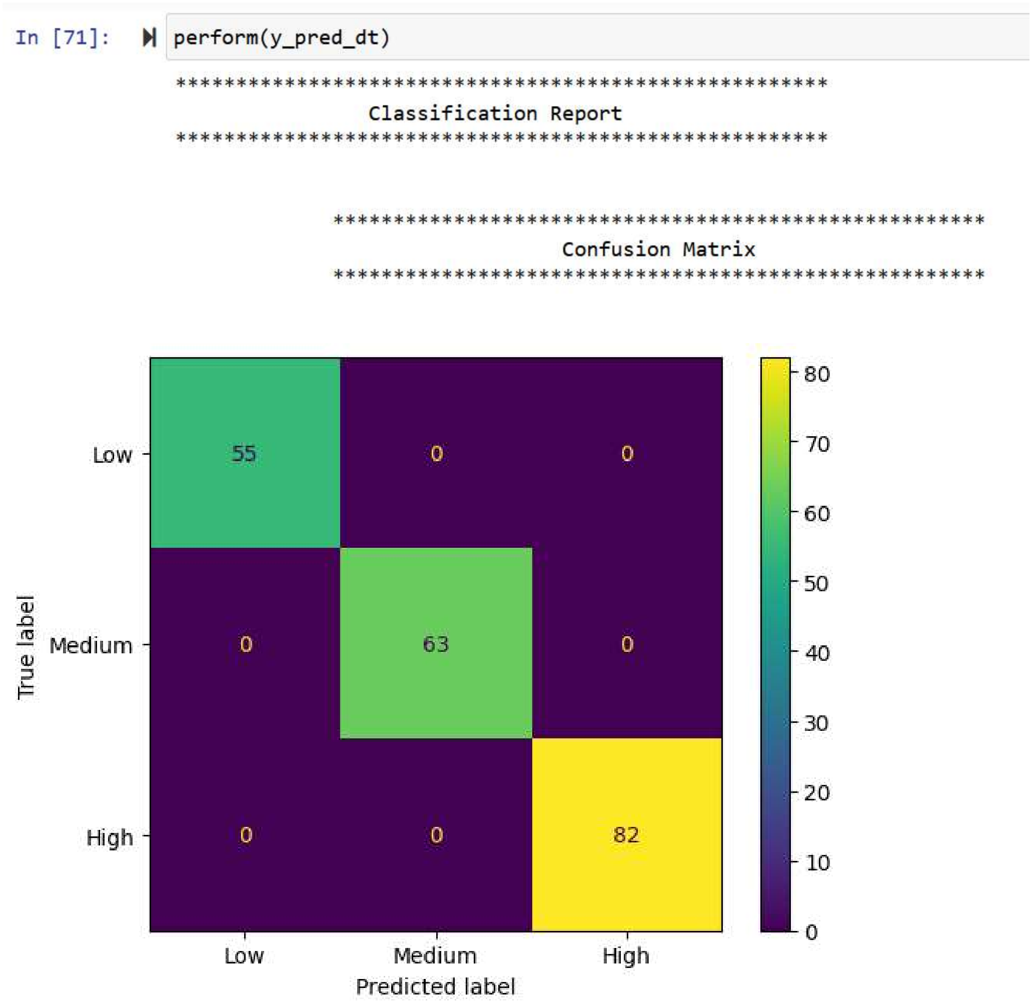
```
In [68]: from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, ConfusionMatrixDisplay, classification_report, precis

In [69]: def perform(y_pred):
             cm = confusion_matrix(y_test, y_pred)
             print("**"*27 + "\n" + " " * 16 + "Classification Report\n" + "**"* 27)
             print("\n\n" + " "* 13 + "**"*27 + "\n" + " "*16 + "\t\tConfusion Matrix\n" + " "*13 + "**"*27)
             cm = ConfusionMatrixDisplay(confusion_matrix = cm, display_labels = ['Low', 'Medium', 'High'])
             cm.plot()
```

13.10 Using ML Algorithm

Here Decision Tree Machine Learning Algorithm is used.

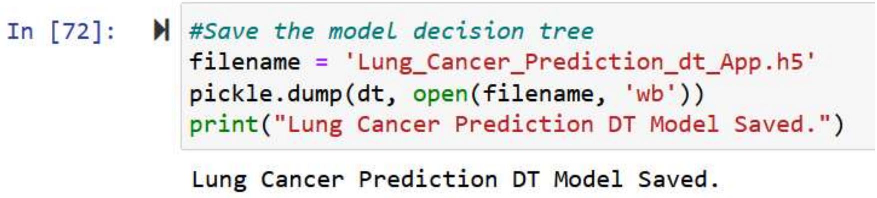


```
In [70]: # Decision Tree Model
         from sklearn.tree import DecisionTreeClassifier
         dt = DecisionTreeClassifier()
         dt.fit(x_train, y_train)
         y_pred_dt = dt.predict(x_test)
         y_pred_dt
```



```
In [71]:    perform(y_pred_dt)

            *******************************************************
                            Classification Report
            *******************************************************


                          *******************************************************
                                          Confusion Matrix
                          *******************************************************
```

13.11 Saving the Model Using Pickel



```
In [72]:    #Save the model decision tree
            filename = 'Lung_Cancer_Prediction_dt_App.h5'
            pickle.dump(dt, open(filename, 'wb'))
            print("Lung Cancer Prediction DT Model Saved.")

            Lung Cancer Prediction DT Model Saved.
```

14. Conclusion

ML models have a long way to go, we can expect ML to replace our local clinicians in the coming decades, and it's pretty exciting! Most models still lack sufficient data and suffer from bias. We can enhance the analysis power of this AI-powered lung cancer detection device by merging CT Scan images of lungs which might help to scan out the small lumps over the lungs; making it provide us with more accurate results. To keep this handy, we can use it in the form of a medical device.

Before integrating it into clinical practice, it is required to validate the accuracy and effectiveness of the lung cancer detection device as it is a matter of human safety. It will be a diagnostic device that might be a boon to all lung cancer detection, leading to better patient outcomes and improved healthcare delivery!

15. References

[1] D. Spiegel, "Cancer", Editor: George Fink, Encyclopedia of Stress (2nd edition), Academic Press, 2007

[2] "Lung Cancer Prediction Using Machine Learning: A Comprehensive Approach", by MA. Jabbar & SA Fathima, IEEE

[3] Konstantina Kourou, Themis P. Exarchos, K.P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, "Machine Learning Applications in Cancer Prognosis and Prediction", Computational and Structural Biotechnology Journal, Volume 13, 2015

[4] "Cancer Prediction Using ML", Ganta Sruthi; Chokkakula Likitha Ram; Malegam Koushik Sai; Bhanu Pratap Singh; Nikhil Majhotra; Neha Sharma, 2022 | 2nd ICIPTM