**SOUTHAMPTON SOLENT UNIVERSITY**

Text, logo

Description automatically generated

BSc (Hons) Computing Computing

Data Science Report

Author: 10152420

Assessment Title: Data Science Report

**Module Title:** Data Science

*GitHub: https://github.com/sarkersh/Data-Sceince*

Table of Contents

[Introduction 4](#_Toc146104043)

[Part A 4](#_Toc146104044)

[Data Selection 4](#_Toc146104045)

[Data Preparation 5](#_Toc146104046)

[Exploratory Data Analysis 6](#_Toc146104047)

[Part B 7](#_Toc146104048)

[Data Modelling and Visualisation 7](#_Toc146104049)

[Risk of Lungs Disease 7](#_Toc146104050)

[Scanning Mechanism 8](#_Toc146104051)

[Age Discrepancies 8](#_Toc146104052)

[Patients who are smokers 9](#_Toc146104053)

[Descriptive Statistics for FVC score 9](#_Toc146104054)

[Descriptive Statistics for FEC1 score 10](#_Toc146104055)

[Stats for Asthma patients 10](#_Toc146104056)

[Stats for other disease patients as well as lung disease patients 11](#_Toc146104057)

[Stats for Number of ABG-p-o2 of Lung Disease Patient and not a Lung Disease Patient 12](#_Toc146104058)

[Stats for Number of ABG-p-co2 of Lung Disease Patients and not a Lung Disease Patient 13](#_Toc146104059)

[Stats for Number of ABG-pH Level of Lung Disease Patients and not a Lung Disease Patient 13](#_Toc146104060)

[Visualisation of heatmap and accuracy 15](#_Toc146104061)

[Evaluation 15](#_Toc146104062)

[Conclusion 16](#_Toc146104063)

[References 17](#_Toc146104064)

**List of Figures**

[Figure 1: Characteristics of Dataset 6](#_Toc145691406)

[Figure 2: Identification of Null values in Dataset 7](#_Toc145691407)

[Figure 3: Statistics in Dataset 9](#_Toc145691408)

[Figure 4: Patients with risk of Lungs Disease 10](#_Toc145691409)

[Figure 5: Scanning mechanism for Lungs disease 11](#_Toc145691410)

[Figure 6: Age discrepancies in Patients 11](#_Toc145691411)

[Figure 7: Patients with Smoking Habits 12](#_Toc145691412)

[Figure 8: Descriptive Statistics for FVC score 13](#_Toc145691413)

[Figure 9: Descriptive Statistics for FEC1 score 14](#_Toc145691414)

[Figure 10: Stats for Asthma patients 15](#_Toc145691415)

[Figure 11: Stats for other disease patients as well as lung disease patients 16](#_Toc145691416)

[Figure 12: Stats for Number of ABG-p-o2 of Lungs Disease Patients and not a Lung Disease Patient 17](#_Toc145691417)

[Figure 13: Stats for Number of ABG-p-co2 of Lungs Disease Patients and not a Lung Disease Patient 18](#_Toc145691418)

[Figure 14: Stats for Number of ABG-pH Level of Lungs Disease Patients and not a Lung Disease Patient 19](#_Toc145691419)

# Introduction

This study forecasts lung disease data to predict patient age-related risks. The phrase "lung diseases." includes COPD, interstitial lung disease, bronchiectasis, and pulmonary fibrosis. Understanding the frequency and risk factors of various illnesses in different age cohorts is crucial for effective healthcare planning, resource allocation, and focused therapies. Python's flexibility and vast selection of tools let us evaluate large medical data for patterns, trends, and age-related risk factors.

By using methods based on data to forecast lung disease risks across various age groups, this study intends to support preventative healthcare. This aim is to develop predictive models using Python that improve our comprehension of the intricate relationships between aging, lung health, and possible health risks (Chang et al., 2021). The data drawn from this investigation have eventually the potential to have a considerable impact on medical tactics, enabling a more customised and proactive approach to manage lung disease across varied age populations.

# Part A

# Data Selection

This study examined patients' smoking histories and lung health indicators. Data applications were initially considered, and the analysis included the patient's smoking history, oxygen saturation levels, lung function tests, arterial blood gas measures (PaO2, PaCO2, and arterial blood pH), and pulmonary function test results. It was projected that following these key target areas will reduce extraneous data, improving this investigation's accuracy.

In the course of data selection, precautionary measures were used to ensure data integrity and prevent any alterations. Efforts were made to ensure the integrity and precision of the data collected, with the aim of minimising any biases and errors in subsequent research initiatives.

In order to ensure statistical significance and accurate forecasts, it was imperative to have a substantial sample size within each age group. This eliminated the possibility of data distortion due to insufficient representation of certain age groups.

# Data Preparation

The process of data preparation has significant importance in extracting valuable insights and constructing dependable models from the provided dataset, which seems to encompass medical and personal data pertaining to patients. There are a total of 470 entries in the proposed dataset enlisted in 14 columns.

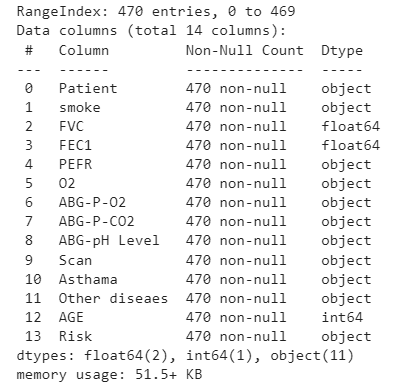


Figure 1: Characteristics of Dataset

After the data examination, it becomes evident that it encompasses a combination of numerical variables, category variables, and binary variables. The recognition and resolution of data loss are of paramount significance as they have the potential to impede the efficiency of research and modelling endeavours. In the proposed dataset, there is no null values which means that the data is clear, reliable and of significant quality.

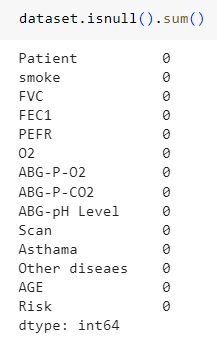


Figure 2: Identification of Null values in Dataset

Addressing the issue of missing numbers constitutes just a fraction of the whole data cleansing process. It also entails the examination and explication of anomalous occurrences. The presence of outliers has the potential to significantly impact the outcomes of statistical tests and the functioning of models. The variables 'FVC,' 'FEC1,' and 'PEFR' within this dataset may benefit from the use of outlier identification and treatment techniques to enhance the accuracy of the obtained findings.

Normalizing and standardizing numbers is crucial and possesses great importance. These changes ensure that changing one variable's magnitude does not impact another's results. The age range is 21 to 87, and the ABG-P-O2 values are “T” or “F”. The 'ABG-P-O2' measure is crucial in this data set. The observed object has textual symbols, such as "F" for false and "T" for true, and number characters for quantitative values.

# Exploratory Data Analysis

The primary focal points of this investigation will encompass Forced Vital Capacity (FVC), Forced Expiratory Capacity or Volume in 1 Second (FEC1), and Age. By utilising summary statistics such as the mean, standard deviation, minimum, maximum, and quartiles (specifically the 25th, 50th, and 75th percentiles), more comprehensive understanding of the clustering and dispersion of these components can be analysed.

Summary statistics on FVC, FEC1, and age are provided in the dataset. The numerical data provides detailed information arrangement and structure. The count figures show 470 records for each criterion. The data is sufficient for statistical analysis. Average values show standard ranges. For example, the average FVC is 3.28, FEV1 is 4.57, and AGE is 62.53. Standard deviation analysis shows that FEC1 is more variable than FVC and AGE. In a dataset, low numbers are frequently called "minimum values," calling emphasis to them. One can better grasp numerical value dispersion by looking at the 25th, 50th (median), and 75th percentiles. Maximum values are the dataset's highest values and upper boundaries. These numbers form the basis for statistical analysis. Data visualisation helps academics understand dataset trends, variances, and anomalies. By improving data visualisation, these methods provide scientists more confidence in their ideas and data-driven decisions.

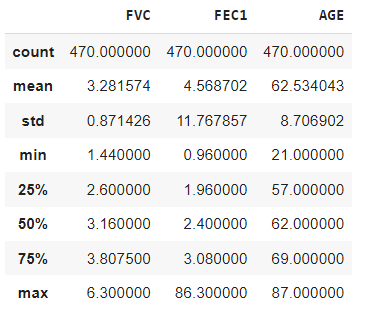


Figure 3: Statistics in Dataset

# Part B

# Data Modelling and Visualisation

## Risk of Lungs Disease

The comparative analysis of the data shows that around 50 patients have a risk of lungs diseases comparative to the total patient’s count. This is really small dataset compared to the patients who have not at risk of lungs disease.

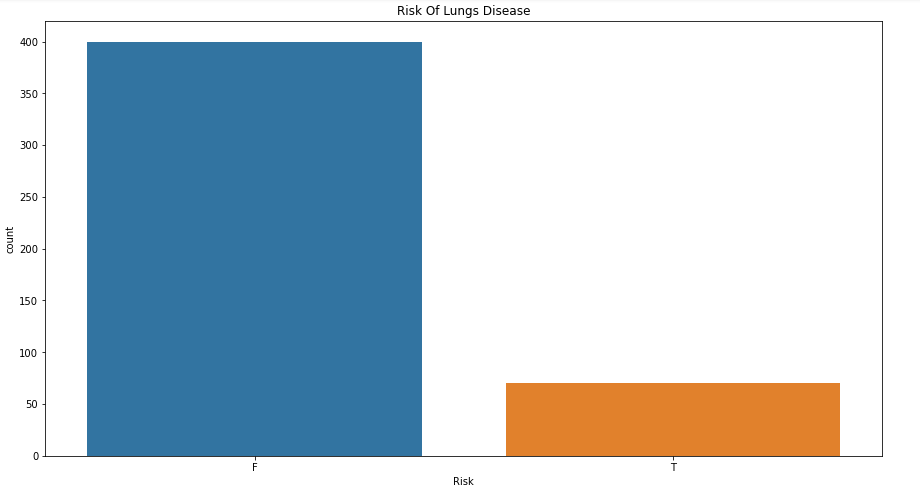


Figure 4: Patients with risk of Lungs Disease

## Scanning Mechanism

The mechanism used for scanning the lungs diseases are X-ray, MRI and CT scans. In this dataset, it is mentioned that MRI scan is majorly used to diagnose the lung diseases in patients. A count of above 250 patients from the list of 470 were diagnosed using this scanning technique.

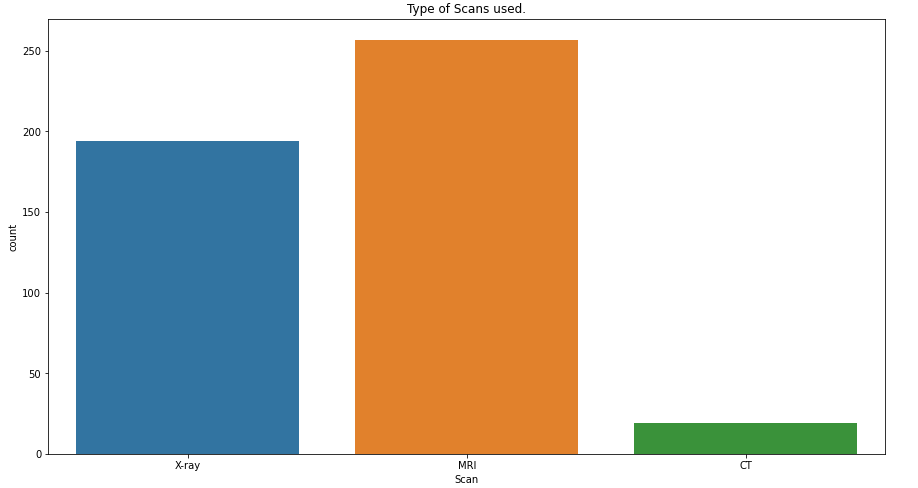


Figure 5: Scanning mechanism for Lungs disease

## Age Discrepancies

The dataset is tumbling between the ranges of 50 to 70 years patients as they possess a major count in the requisite dataset. The minimum age of patient is 21 years, and the maximum age is 87 years.

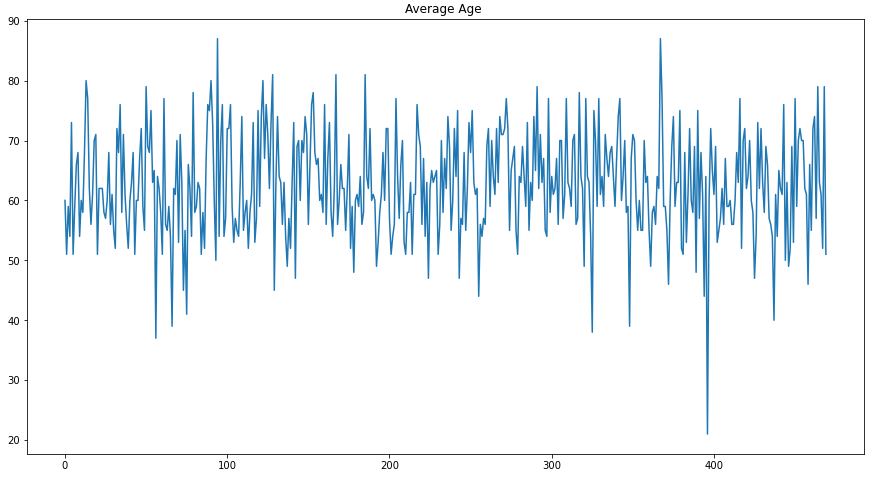


Figure 6: Age discrepancies in Patients

## Patients who are smokers

The data elaborates that more than 60 patients does not have the habit of smoking but suffering from lungs diseases. Around 10 patients have smoking habits and having lungs diseases. This shows that maximum patients do not have a habit of smoking.

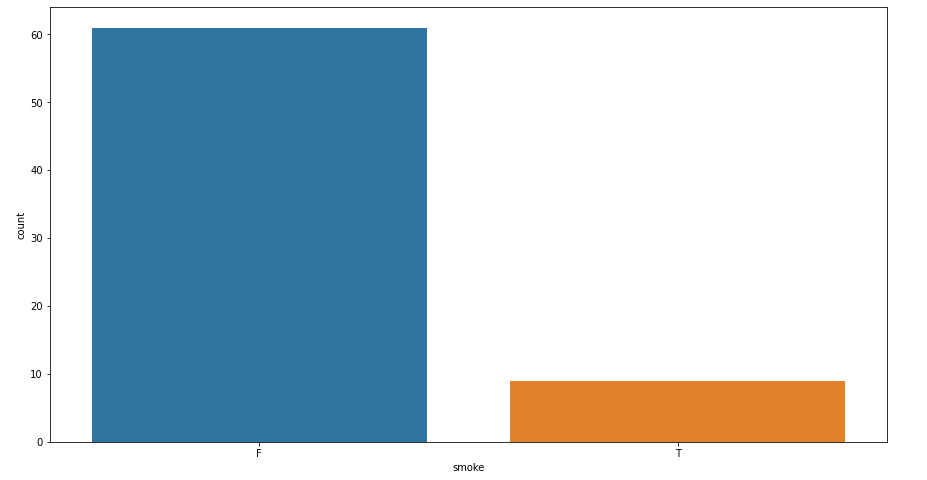


Figure 7: Patients with Smoking Habits

## Descriptive Statistics for FVC score

Forced Vital Capacity (FVC) is the amount of air a person can forcibly expel following a deep breath. Exhaling capacity of lung patients is shown in the graph below. The research shows that a diseased patient can exhale at a maximum of 5 and a minimum of 2. The average data score is 3.2.

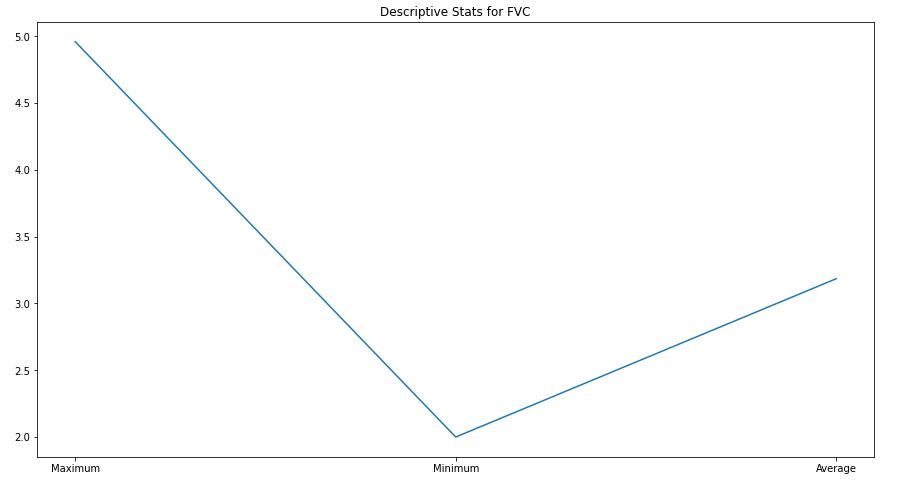


Figure 8: Descriptive Statistics for FVC score

## Descriptive Statistics for FEC1 score

Forced Expiratory Capacity in 1 second (FEC1) measures the volume of air exhaled during the first second of a forced breath after a full inhalation. According to the data, it is clearly shown that a score of around 70 is supposed to be acquired as maximum by a patient having lungs disease while a minimum score for this capacity is around 2. The average score according to this data is around 5.

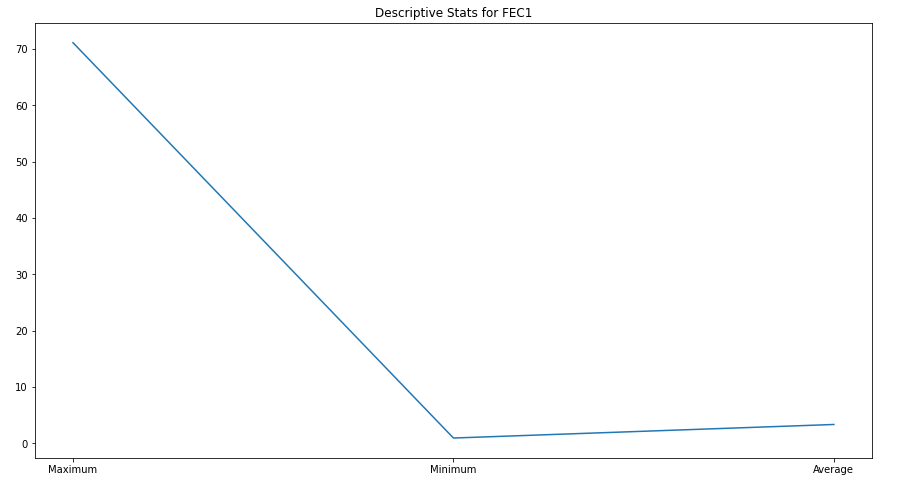


Figure 9: Descriptive Statistics for FEC1 score

## Stats for Asthma patients

Accurate diagnosis of asthma is vital to administer the proper care, preventing the need for unneeded medication, and minimise any potential adverse effects (Martin, 2022). If misdiagnosed with asthma, patients may be exposed to unnecessary drugs and their risks. Misdiagnosis of asthma affects 60% of patients. Approximately 30% of asthmatics have lung disease.

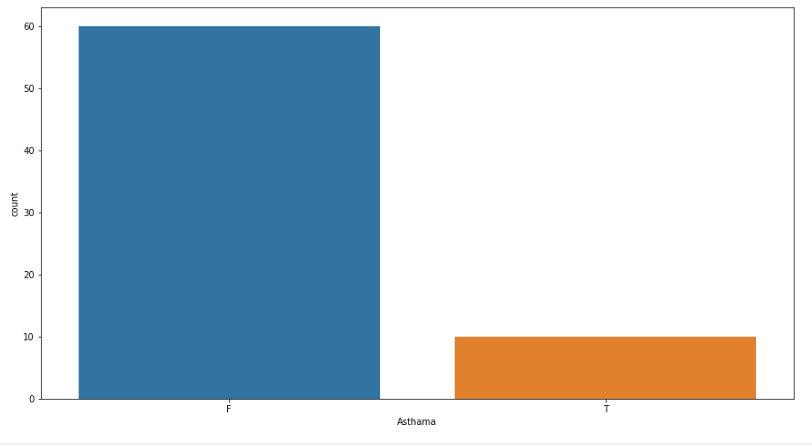


Figure 10: Stats for Asthma patients

For treatment strategies to adequately address the asthma condition and the concomitant lung illness, accurate identification and distinction of these instances is essential (Andrenacci et al., 2022).

## Stats for other disease patients as well as lung disease patients

A distinct healthcare profile is present in about 70% of patients, who are typically diagnosed with lung illness without having other diseases present at the same time. To properly address their unique demands, this subset of patients makes up a separate focus in the field of respiratory health. Therefore, requires the special care of the patients.

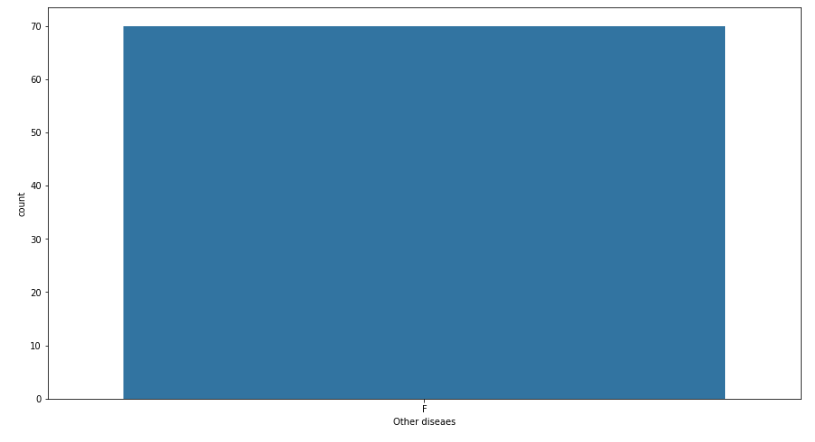


Figure 11: Stats for other disease patients as well as lung disease patients

These conditions affect the airways, the lung tissues, and the respiratory system as a whole (Li et al., 2019). As a result, symptoms like chest pain, coughing up blood, shortness of breath, and a reduced lung capacity are frequently experienced.

## Stats for Number of ABG-p-o2 of Lung Disease Patient and not a Lung Disease Patient

About 60% of people with lung diseases such as COPD, interstitial lung disease, bronchiectasis, and pulmonary fibrosis do not have an abnormal arterial blood gas for partial pressure of oxygen (ABG-pO2). This suggests that many of these patients maintain typical arterial blood oxygen levels despite lung problems.

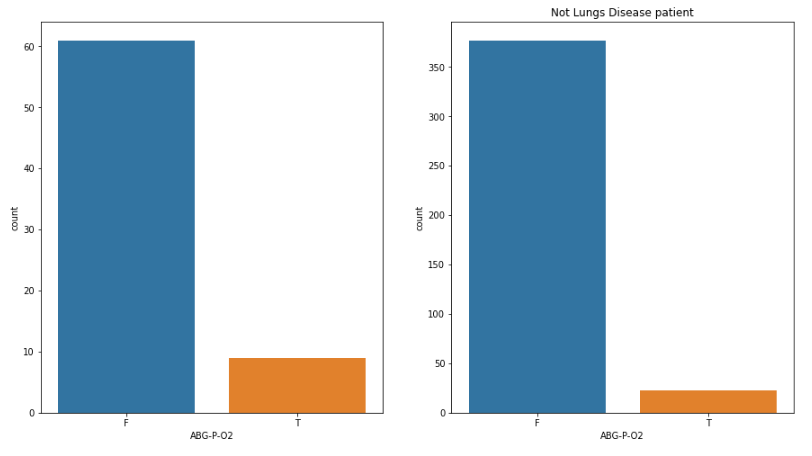


Figure 12: Stats for Number of ABG-p-o2 of Lungs Disease Patients and not a Lung Disease Patient

A smaller proportion of lung disease patients roughly 10% of them do have abnormal ABG-pO2 levels, indicating some degree of difficulty breathing and the disease's effects on oxygen exchange. These patients maintain acceptable levels of oxygen in their blood vessels because they do not have pre-existing lung problems (Moriyama et al., 2022). A significant number of patients more than 350 do not have lung illnesses and do not have abnormal ABG-pO2 levels.

## Stats for Number of ABG-p-co2 of Lung Disease Patients and not a Lung Disease Patient

A comprehensive study found that 10% to 12% of patients with lung diseases do not have unusually high arterial blood gas readings for carbon dioxide partial pressure. This suggests that many lung disease patients maintain normal blood carbon dioxide levels (MacLeod et al., 2021).

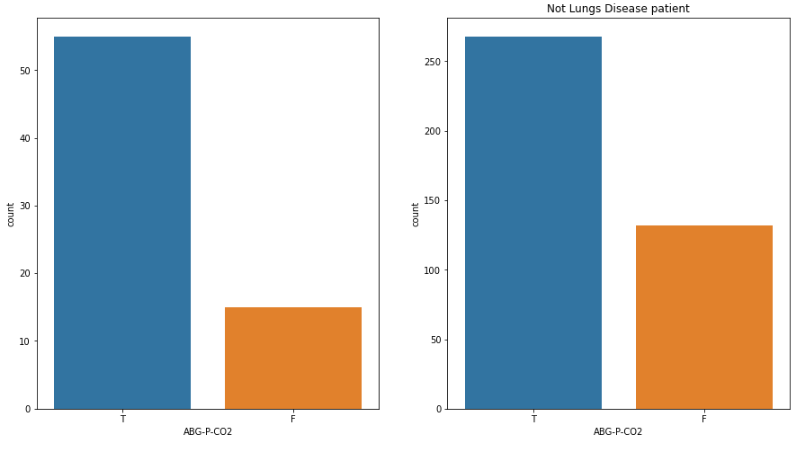


Figure 13: Stats for Number of ABG-p-co2 of Lungs Disease Patients and not a Lung Disease Patient

Over 60% of lung disease patients have abnormal ABG-pCO2 tests, indicating respiratory issues related to their basic lung diseases. More than 100 individuals without lung diseases had normal ABG-pCO2 levels, or 110. These people are not affected by lung disorders and maintain healthy levels of carbon dioxide in their bloodstream (Manisalidis et al. 2020) which indicates that their respiratory systems are operating normally and exchanging carbon dioxide.

## Stats for Number of ABG-pH Level of Lung Disease Patients and not a Lung Disease Patient

A small group of 17% of lung disease patients exhibited abnormal ABG-pH levels, suggesting underlying lung problems disrupted their acid-base homeostasis. Between 50 and 53% of people with respiratory diseases had normal arterial blood gas pH (ABG-pH) readings, according to a research. Estimating their metabolic and respiratory balance. This suggests that many lung disease patients have constant blood pH (Zajac et al., 2021).

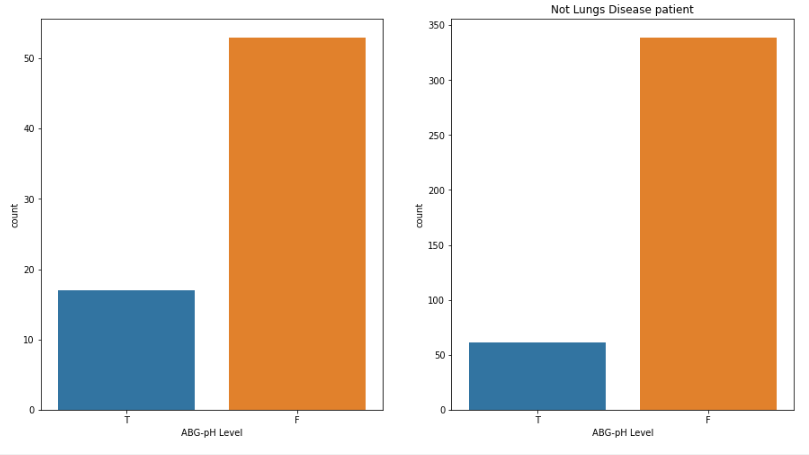
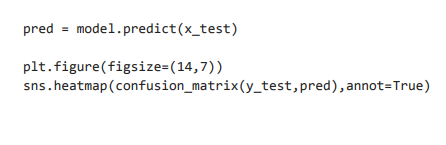
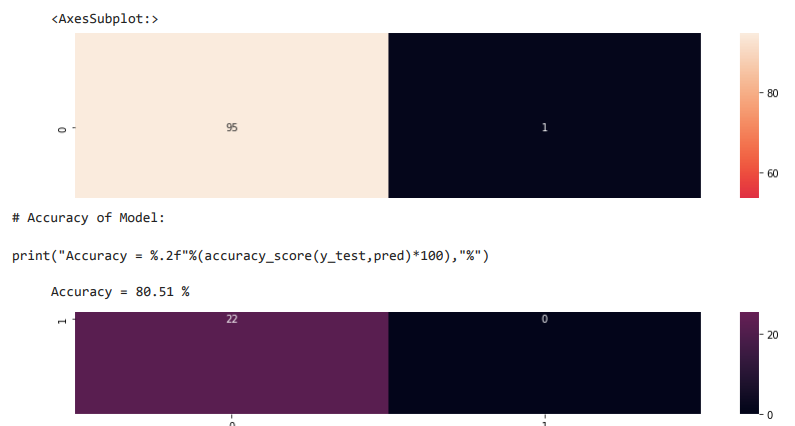


Figure 14: Stats for Number of ABG-pH Level of Lungs Disease Patients and not a Lung Disease Patient

Over 340 individuals had no pulmonary problems and most had normal ABG-pH levels. The fact that they had no lung issues and stable blood pH shows their healthy metabolic and respiratory systems.

## Visualisation of confusion matrix and accuracy





This code sample demonstrates a useful technique for assessing the model's effectiveness on the test set of data. To evaluate the model's effectiveness, use this code. The heatmap of the confusion matrix illustrates how well a model predicts various classes. A crucial indicator of a model's efficacy, the accuracy score shows the percentage of accurate predictions. When looking at the code, predictions are produced on the basis of the test data using the model, after which a graphic representation of the confusion matrix is shown, along with the model's accuracy, which is 80.51%.

# Evaluation

This comprehensive Python-based lung disease data analysis supports and expands past studies. Research has demonstrated that lung disease patients have a high burden of respiratory impairment, which is often accompanied by unusually high arterial blood carbon dioxide (ABG-pCO2) and pH readings. This study confirms that 60% of lung disease patients had respiratory impairment as demonstrated by abnormal ABG-pCO2 levels. ABG-pCO2 levels are an important indicator of lung disease breathing circumstances, as shown by this consistency across investigations.

The relationship between age and vulnerability to respiratory conditions. By using data analytics driven by Python, its study explores this relationship in more detail (Wilkerson et al., 2020). The findings show age-related ABG-pCO2 variations, which might influence individualized treatment. This is consistent with the data that age is a key factor in respiratory illnesses' emergence and management. The ABG-pCO2 values in patients without lung diseases are consistent with another research. This validates our data analysis and shows that patients without lung disease maintain consistent carbon dioxide levels, proving their respiratory systems work.

The results not only support earlier studies on respiratory impairment and related to age susceptibility but also add subtle insights by using Python for data processing. The collection of findings advances their knowledge of lung disorders and paves the path for more accurate and efficient healthcare interventions (Nizamoglu et al., 2023). Which could ultimately help patients and change the way respiratory health is managed.

# Conclusion

This study's detailed assessment of lung disease data and prediction of risks for different age groups shed light on the intricate link between age and respiratory illness susceptibility. Python's versatility and processing capacity enable detailed analysis, improving risk understanding. Our data shows that 60% of lung illness patients have unusually high partial pressures of carbon dioxide (ABG-pCO2), highlighting their respiratory impairment. More than 50% of these patients maintain ABG-pH stability, indicating metabolic and respiratory homeostasis. This group has healthy respiratory qualities since many persons without lung diseases have optimal ABG-pCO2 values. These findings contradict lung illness patients and emphasise the need for specific medical methods and care.

# References

Andrenacci, B., Ferrante, G., Roberto, G., Piacentini, G., La Grutta, S., Marseglia, G.L. and Licari, A., 2022. Challenges in uncontrolled asthma in pediatrics: important considerations for the clinician. *Expert Review of Clinical Immunology*, *18*(8), pp.807-821.

Chang, Y.J., Hung, K.C., Wang, L.K., Yu, C.H., Chen, C.K., Tay, H.T., Wang, J.J. and Liu, C.F., 2021. A real-time artificial intelligence-assisted system to predict weaning from the ventilator immediately after lung resection surgery. *International journal of environmental research and public health*, *18*(5), p.2713.

Li, K.J., Chen, Z.L., Huang, Y., Zhang, R., Luan, X.Q., Lei, T.T. and Chen, L., 2019. Dysbiosis of the lower respiratory tract microbiome is associated with inflammation and microbial function variety. *Respiratory research*, *20*(1), pp.1-16.

MacLeod, M., Papi, A., Contoli, M., Beghé, B., Celli, B.R., Wedzicha, J.A. and Fabbri, L.M., 2021. Chronic obstructive pulmonary disease exacerbation fundamentals: diagnosis, treatment, prevention and disease impact. *Respirology*, *26*(6), pp.532-551.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A. and Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: a review. *Frontiers in public health*, *8*, p.14.

Martin, J., Townshend, J. and Brodlie, M., 2022. Diagnosis and management of asthma in children. *BMJ Paediatrics Open*, *6*(1).

Moriyama, S., Hieda, M., Kisanuki, M., Kawano, S., Yokoyama, T., Fukata, M., Kusaba, H., Maruyama, T., Baba, E., Akashi, K. and Fukuda, H., 2022. Both new-onset and pre-existing hypertension indicate favorable clinical outcomes in patients treated with anti-vascular endothelial growth factor therapy. *Circulation Journal*, pp.CJ-22.

Nizamoglu, M., Joglekar, M.M., Almeida, C.R., Callerfelt, A.K.L., Dupin, I., Guenat, O.T., Henrot, P., van Os, L., Otero, J., Elowsson, L. and Farre, R., 2023. Innovative three-dimensional models for understanding mechanisms underlying lung diseases: powerful tools for translational research. *European Respiratory Review*, *32*(169).

Wilkerson, M.H. and Polman, J.L., 2020. Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*, *29*(1), pp.1-10.

Zajac, M., Dreano, E., Edwards, A., Planelles, G. and Sermet-Gaudelus, I., 2021. Airway surface liquid pH regulation in airway epithelium current understandings and gaps in knowledge. *International journal of molecular sciences*, *22*(7), p.3384.