EXPLORING THE RELATIONSHIP BETWEEN AIR QUALITY INDEX AND LUNG CANCER MORTALITY IN UNITED STATE : PREDICTIVE MODELING AND IMPACT ASSESSMENT

BY

MD. SAMIUL ISLAM
ID: 0242310005103026

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By
Dr. Md Zahid Hasan

Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By
Mr. Abdus Sattar

Assistant Professor and Coordinator M.Sc
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JULY 2024

APPROVAL

This Thesis titled "**Exploring the Relationship Between Air Quality Index and Lung Cancer Mortality in United State: Predictive Modeling and Impact Assessment.**„ submitted by **Md. Samiul Islam**, ID No: **0242310005103026** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 06-07-2024.

BOARD OF EXAMINERS

Chairman

Narayan Ranjan Chakraborty
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Ohidujjaman, PhD
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Sadat Hossain
Data Scientist
Risk Management Division,
BRAC Bank Limited

DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Dr. Md Zahid Hasan, Associate Professor, Department of CSE,** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Md Zahid Hasan

Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Abdus Sattar
Assistant Professor & Coordinator M.Sc
Department of CSE
Daffodil International University

Submitted by:

Md Samiul Islam

ID: 0242310005103026
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Dr. Md Zahid Hasan**, Associate Professor, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Data Mining" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor and Head, Department of CSE**, for his kind help to finish our thesis and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

In order to understand the effects of air pollution on health, we use advanced predictive modeling to examine the relationship between the Air Quality Index (AQI) and lung cancer mortality across major urban areas in the United States. We concentrate on important pollutants that are known to contribute to poor air quality, like PM2.5 and NO2. We use statistical analysis and machine learning techniques, such as linear regression and XGBoost, to accurately predict health outcomes through an extensive dataset analysis, which includes historical air quality readings from the Environmental Protection Agency (EPA) and health records from the Centers for Disease Control and Prevention (CDC). Our results show a strong correlation between increased lung cancer mortality and high levels of particulate pollutants. This relationship emphasizes how important it is to implement focused public health initiatives to lower exposure to these dangerous contaminants. In order to mitigate the negative health effects of air pollution, we support legal changes based on our findings, such as stronger standards for air quality and improved mechanisms for monitoring pollution.This study not only establishes the foundation for future research examining more comprehensive environmental and public health policies, but it also gives vital information about the risks associated with air pollution to U.S. lawmakers and health professionals. By demonstrating the direct relationship between air quality and healthcare outcomes, Our study is an essential resource for enhancing the dynamics of urban health, especially in the quickly growing metropolitan areas of the United States.

Keywords: Quality Index, Random Forest, XGBoost, lawmakers, mortality, contaminants, correlation, lower exposure.

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2: Exploratory Data Analysis	15
Figure 3.3: Exploratory Data Analysis ploting	16
Figure 3.4 : Different kind of predictor variables	16
Figure 3.5 : Pearson correlation coefficient matrix	18
Figure 3.6 : Spearman correlation coefficient matrix	19
Figure 3.7 : Lung cancer mortality distribution by AQI with T-test	20
Figure 3.8 : Chi-Square test for lung cancer mortality by AQI category	21
Figure 3.9 : Error Rate vs Number of Neighbors for KNN	23
Figure 3.10 : Error Rate vs Number of Neighbors for Decision Tree.	26
Figure 3.11 : Mean Absolute Error for Different models.	31
Figure 3.12 : Root Mean Squared Error for Different Models	32
Figure 3.13 : R-squared for Different Models.	32
Figure 4.1: Correlation matrix heatmap of AQI components and lung cancer mortality.	37
Figure 4.2 : Cross-validation results for selected models.	39
Figure 4.3 : Actual vs. predicted lung cancer mortality rates.	39

LIST OF TABLES

TABLES	PAGE NO
Table 2.1 : Air Quality Index (AQI) for United state	6
Table 4.1: Shows the accuracy results of different machine learning models	35
Table 4.2: Descriptive statistics of key variables	36
Table 4.3: Correlation matrix of AQI components and lung cancer mortality	36
Table 4.4: Results of hypothesis tests	37
Table 4.5 : Performance metrics for various models	38
Table 4.6 : Evaluation metrics for the best-performing model	39

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	ii
Declaration	iii
Acknowledgement	iv
Abstract	v
List of Figure	vi
List of table	vii
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-4**
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	3
1.5 Objectives	3
1.6 Report Layout	4
CHAPTER 2: BACKGROUND	**5-11**
2.1 Terminologies	5
2.2 Related Works	6
2.3 Comparative Analysis and Summary	9
2.4 Scope of the Problem	10
2.5 Challenges	11
CHAPTER 3: RESEARCH METHODOLOGY	**11-32**
3.1 Research Subject and Instrumentation	11
3.2 Data Collection Procedure	13

3.3 Data Cleaning and Preprocessing 13

3.4 Statistical Analysis 14

3.5 Machine Learning Models and Predictive Analysis 21

3.5.1 K-Nearest Neighbors (KNN) 21

3.5.2 Logistic Regression 23

3.5.3 Decision Tree 25

3.5.4 Random Forest 27

3.5.5 Artificial Neural Networks 28

3.5.6 XGBoost 30

3.6 Model Training and Validation 31

3.7 Implementation Requirements 32

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **34-40**

4.1 Experimental Setup 34

4.2 Experimental Results Analysis 34

4.3 Model Performance and Evaluation 38

4.4 Discussion 40

CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH **41-42**

5.1 Summary of the Finding 41

5.2 Conclusions 41

5.3 Implication for Further Research 41

5.4 Recommendations 42

5.5 Final Thoughts 42

REFERENCES **43-44**

CHAPTER 1
Introduction

1.1 Introduction

Worldwide, air pollution poses a serious threat to human health and the environment. There is strong scientific evidence connecting air pollution exposure to lung cancer and other health problems, such as respiratory conditions. An uniform measure of air pollution levels and their effects on health is the Air Quality Index (AQI). Because of their large populations, heavy traffic, and extensive industrial activity, American cities like Los Angeles and New York are especially vulnerable to air pollution[1].

The United States has encountered serious problems with air quality. For example, in 2021, the Environmental Protection Agency (EPA) classified many cities as "unhealthy" due to AQI values regularly above 150. Pollutants such as PM2.5 can accumulate in such environments to concentrations that are many times higher than WHO-recommended thresholds. The increasing number of lung cancer occurrences in the country is mostly due to the serious health hazards posed by these pollution levels[2][4].

Cities in America frequently have high global pollution indices, highlighting a persistent problem with the environment and public health. For instance, several days in 2021 had 'Unhealthy' air quality levels in places like Los Angeles, where PM2.5 concentrations were abnormally high, increasing the hazards to millions of people's health.

This study uses sophisticated predictive modeling approaches to investigate the relationship between the American lung cancer mortality rate and the AQI. We concentrate on important pollutants that cause poor air quality, such as nitrogen dioxide (NO2), ozone (O3), PM2.5, and PM10. [3] To achieve high predicted accuracy, a comprehensive dataset comprising historical health records and air quality measurements is examined using both contemporary machine learning algorithms, like KNN, Logistic Regression, Decision tree, Random Forest, XGBoost and ANN conventional statistical methods.

Our results show a robust correlation between high pollution levels and higher lung cancer death rates, especially in areas where PM2.5 concentrations are higher than 100 $\mu g/m^3$. These findings emphasize how urgently effective public health initiatives to lower exposure to dangerous chemicals are needed.

In order to lessen the negative health effects of air pollution, we advise enacting stronger air quality standards and enhancing pollution monitoring systems. In addition to educating American lawmakers and medical professionals on the risks posed by air pollution, this study lays the foundation for future investigations into more comprehensive environmental and public health initiatives. This research offers a critical tool for improving urban health dynamics and directing public health interventions in rapidly expanding American cities by illuminating the direct relationship between air quality and health outcomes.

1.2 Motivation

The substantial rise in air pollution levels in the biggest cities in the United States and the associated rise in lung cancer cases are the driving forces behind this investigation. According to recent research, there is a worrying correlation between higher levels of air pollution, specifically PM2.5 and NO2, and higher mortality rates from lung cancer. It is essential to comprehend the relationship between lung cancer mortality and the Air Quality Index (AQI) in order to influence public health policy and spur the creation of successful interventions. These kinds of measures are necessary to guarantee public health and safety, lessen the negative health effects of air pollution, and provide direction for industrial and urban development rules.[5] This study is to close the knowledge gaps in the field and offer solid facts to assist in the creation of policies targeted at lowering air pollution and its adverse health effects.

1.3 Rationale of the Study

The purpose of this study is to offer verifiable proof of the connection between air quality index and lung cancer mortality. By emphasising the direct effects of pollutants like PM2.5 and NO2, we hope to draw attention to how urgently effective public health actions are needed. The goal of this research is to provide lawmakers, health care providers, and local government agencies with the information they need to develop workable plans for lowering air pollution. In the end, this study aims to address a major public health concern with practical solutions in order to assure a higher standard of living and promote community health.

1.4 Research Questions

Several important concerns that directly impact the well-being of communities across the United States serve as the foundation for our research:

i. What is the relationship between the Air Quality Index (AQI) and the varying lung cancer death rates throughout the climatic zones of the United States, such as those found in rural and industrial areas?
ii. Which specific pollutants, such fine particulate matter (PM2.5) or nitrogen dioxide (NO2), are most significantly associated with an increased risk of lung cancer mortality in urban environments?
iii. How much can the AQI changes of the last 10 years in major U.S. cities be used to predict the death rate from lung cancer using machine learning models like ANN, KNN, Random Forest, Decision Tree, and XGBoost?
iv. Based on our research, what focused public health initiatives have been shown to be successful in areas with historically low air quality, and what fresh approaches can be put into practice?

1.5 Objectives

The study's results are intended to offer practical guidance and encourage important developments in American public health regulations pertaining to air quality:

i. A detailed investigation that outlines the correlation between the AQI and the mortality rate from lung cancer in each of the 47 states. This investigation will assist in identifying the disparities in the health effects of air quality in urban and rural locations, as well as regional vulnerabilities.
ii. Determination and measurement of particular pollutants that raise the risk of lung cancer considerably. In order to provide precise targets for mitigation activities, this output will include a thorough analysis of the impacts of important pollutants, such as PM2.5 and NO2, in densely populated metropolitan areas.
iii. The development and validation of sophisticated prediction models that make use of previous AQI data to accurately anticipate the mortality rate from lung cancer. These models, which provide useful resources for public health forecasting and planning, will be assessed for accuracy and dependability.

iv. Policy suggestions are formulated using modeling insights and empirical data. This will cover a number of suggested interventions that have been efficaciously modeled, providing solutions that are scalable for lowering air pollution and lessening its negative health effects. The study will also offer fresh, creative approaches to public health that are based on the most recent findings.

1.6 Report layout:

The study report is divided into six chapters, each methodically investigating how the Air Quality Index (AQI) affects the mortality rate from lung cancer in the US:

- **Chapter 1:** Introduction, providing an overview, motivation, and rationale for the research, along with research questions and objectives.
- **Chapter 2:** Literature review, examining relevant material, theoretical framework, and comparative analysis.
- **Chapter 3:** Methodology, detailing data gathering methods, data cleaning, preprocessing, and analysis tools.
- **Chapter 4:** Experimental results and discussion, presenting findings and placing them in the context of previous research and study hypotheses.
- **Chapter 5:** Conclusions and recommendations, exploring implications for public health and air quality regulations, and identifying areas for further study.
- **Chapter 6:** Summary of main findings, general research conclusions, closing thoughts, and references and appendices.

CHAPTER 2

Background

2.1 Terminologies

It is crucial to clarify important terms and concepts in order to provide readers with an accurate understanding of the study:

Air Quality Index (AQI): This device is used all around the country to show the level of pollution in the air or cleanliness on any particular day. Higher scores on the AQI scale, which ranges from 0 to 500, indicate worse air quality and more possible health hazards.

Lung Cancer: This is a term used to describe a cancerous tumour that starts in lung tissue. Smoking is still the most well-known cause, but air pollution is becoming more and more acknowledged as a risk factor, especially in cities.

PM2.5 and PM10: Particulate matter small enough to get into the tiniest passageways of the respiratory system is referred to by these names. PM2.5 and PM10 particles have dimensions of smaller than 2.5 and 10 micrometres, respectively. These two kinds of particles can come from a number of sources, such as industrial operations and emissions from vehicles, and are significant hazards in urban areas.

Common Air Pollutants: Among the items in this category are:

Nitrogen dioxide (NO2): which is frequently produced by burning fossil fuels in vehicles and other operations.

Sulphur dioxide (SO2): Mainly created when power stations and other industrial facilities burn fossil fuels.

Carbon monoxide (CO): A colourless, odourless gas that is frequently found in vehicle exhaust and is a consequence of incomplete combustion.

Ozone (O3): A gas produced in the atmosphere by the reaction of sunlight with contaminants like nitrogen oxides.

Predictive Modeling: In order for predicting future events, statistical techniques are applied to previous data. In the framework of our research, we apply predictive modeling to project lung cancer incidences in the future based on changes in air quality indexes.

These definitions serve as the basis for our analysis and discussion, making the complicated connections between air quality and health outcomes that we investigate in this study easier to understand.

2.2 Related Works

Analysing air pollution has become crucial for comprehending and recording pollution levels across different places. The prediction, forecasting, and control of pollution levels depend heavily on machine learning techniques. In this research, three machine learning algorithms—Support Vector Machine (SVM), M5P model tree, and Artificial Neural Network (ANN)—that create forecasting models for ground-level ozone, nitrogen dioxide, and sulphur dioxide are presented. When various features are added to multivariate modelling, the M5P algorithm outperforms other models in terms of forecasting performance[11]. ANN, gradient boost, and LUR machine learning algorithms were used to real-time sensor data from downtown Toronto, Canada, in order to forecast. Microaethalometers were used at GPS locations to gather PM2.5 and BC data. The microaethalometer was used to collect the data. According to the study, ANN and the gradient boost technique were more accurate for large datasets, whereas LUR performed better with smaller datasets[12]. Five criterion pollutants are used to calculate United State's Air Quality Index (AQI): PM2.5, PM10, NO2, CO, SO2, and O3. The Air Quality Index indicates the degree of fresh or polluted air as well as the health effects. The following will be the AQI's characteristics: a single number between 0 and 500 that is calculated using the measured concentrations of pollutants in a certain area [13].

Table 2.1 : Air Quality Index (AQI) for United state.

AQI Range	Category (English)	Category (Bangla)	Health effects
0-50	Good	ভালো	Good air quality
51-100	Moderate	মোটামুটি	Low air pollution and no ill effects on health.

101-200	Caution	সতর্কতামূলক	Moderate pollution, Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Unhealthy	অস্বাস্থ্যকর	Mild aggravation of symptoms among high-risk persons, like those with heart or lung disease
301-400	Very Unhealthy	খুব অস্বাস্থ্যকর	Significant aggravation of symptoms and decreased exercise tolerance in persons with heart or lung disease.
401-500 or more	Extremely Unhealthy	অত্যন্ত অস্বাস্থ্যকর	Severe aggravation of symptoms and endangers health

The work conducted by Sedef Cakir and Moro Sita provides particularly useful insights into the dynamics of air pollution. They used Artificial Neural Networks (ANNs) and Multiple Linear Regression (MLR) in their research, which was carried out in Nicosia, to estimate the amounts of important pollutants like NO2, O3, and PM10. They discovered that ANNs frequently outperformed MLR models, particularly in situations involving complicated environmental data, because of their capacity to simulate complex, non-linear connections. My research, which focuses on air quality indices and lung cancer rates in metropolitan U.S. settings, is directly related to this finding, which highlights the potential of sophisticated machine learning approaches in environmental health studies. My work attempts to improve the accuracy of health outcome forecasts associated to air pollution by integrating strong datasets and similar predictive models. It does this by drawing on established approaches to help public health initiatives[7].

Machine learning is improving our understanding of air quality and its effects on health. In their groundbreaking work, Suresh Kumar Natarajan and colleagues refined decision tree algorithms using Grey Wolf Optimisation, which greatly increased the accuracy of AQI forecasts in India's busy metropolises, such as Hyderabad and Delhi. Their creative strategy shows how combining machine learning and optimisation approaches may provide more accurate forecasts; I hope to duplicate and modify this approach to learn more about how air quality affects lung cancer mortality in metropolitan U.S. settings. The promise of

sophisticated computational tools to forecast and alleviate health risks associated with poor air quality is reinforced by this research, which not only serves as an example of the advancements in environmental health analytics but also serves as inspiration for my methodology[8].

Around the world, research is concentrated on how to combine different machine learning approaches to manage the complex patterns of data on air pollution. These research use ensemble approaches, neural networks, and support vector machines to improve the accuracy of Air Quality Index (AQI) prediction models. When it comes to handling the unexpected unpredictability of environmental data, ordinary linear statistical models are usually insufficient, but more sophisticated methods are showing to be more successful. This indicates a dramatic change in the direction of the adoption of more advanced systems capable of properly interpreting complicated air quality data and dynamically adapting to it[9].

The study presented in [10] provides a fascinating investigation into improving air quality predictions using cutting-edge machine learning methods. The work not only increases forecast accuracy across multiple major Indian cities by merging Grey Wolf Optimisation with Decision Tree methods, but it also establishes a standard for mixing several computational methodologies to manage complicated environmental data. The effectiveness of this model, demonstrated by comprehensive performance measures like RMSE and R-square, highlights the potential of advanced algorithms to greatly enhance evaluations of urban air quality, making it a useful benchmark for related environmental research.

The research presents an advanced prediction model for air quality forecasting, which is essential given the harmful effects of pollutants like PM2.5, which are known to enter the lungs deeply. This model estimates air quality up to 48 hours ahead of time using a combination of Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks, in contrast to standard methods that only offer real-time data. Through the integration of many data sources, such as terrain influences and historical meteorological data, this novel technique improves forecast accuracy. After being validated using datasets from Beijing and Taiwan, the model performs better than previous approaches and provides a reliable tool for proactive public health management and air quality policy-making[14].

The article "Air Pollutants and their Possible Health Effects at Different Locations in Dhaka City" examines the harmful impacts that different air pollutants have on the city's population's health. According to studies, exposure to pollutants such VOCs, CO, NOx, and PM2.5 is linked to a number of major health problems, including lung cancer and respiratory and cardiovascular disorders. In order to lessen the negative effects of air pollution on Dhaka's urban population, the study emphasises the critical necessity for routine air quality monitoring and efficient management techniques. The results highlight how crucial it is to address air quality concerns in order to promote public health safety and sustainable urban life[15].

The "Ambient Air Quality in Bangladesh 2012-2018" publication provides an overview of the comprehensive air quality monitoring programmes carried out in eight major Bangladeshi cities as part of the Clean Air and Sustainable Environment (CASE) project. It describes how data on air quality were compiled and analysed between 2012 and 2018, with an emphasis on pollutants including ozone (O3), sulphur dioxide (SO2), nitrogen oxides (NOx), and particulate matter (PM2.5 and PM10). The paper highlights the serious problems with air pollution caused by high particulate matter levels, especially in the dry season. It covers the use of tactics for managing air quality, such as trend analysis of air quality, the effects of different contaminants on public health, and the efficacy of regulatory controls. The publication also contains thorough explanations of the approaches that were employed[16].

2.3 Comparative Analysis and Summary

The purpose of this study is to close the knowledge gap about the relationship between lung cancer mortality and air pollution in the particular context of the United States. Although prior studies have demonstrated the detrimental effects of pollutants such as PM2.5, PM10, Ozone (O3), and NO2 on lung cancer incidence, it is necessary to investigate these effects in light of the industrial emissions and automobile pollution that occur in the United States. By concentrating on these variables, our research adds to a more thorough comprehension of the ways in which certain contaminants affect lung cancer mortality in the United States, offering knowledge that may direct successful public health initiatives.

2.4 Scope of the Problem

One major and complex public health concern in the United States is the link between lung cancer mortality and air quality. In order to guarantee a comprehensive investigation of how air pollution affects lung cancer outcomes across the country, this section summarizes the geographical, temporal, and health-related components of the problem.

2.4.1 Geographic Scope

This research includes all of the United States, with an emphasis on regions with considerable variations in air quality and population density, such as industrial zones (like Pittsburgh, Cleveland), metropolitan centres (like Los Angeles, New York City), and rural areas. Our goal in analysing these various places is to comprehend how various environments influence changes in air quality and how those variations affect the death rate from lung cancer.

2.4.2 Temporal scope

Long-term changes in air quality and health outcomes may be thoroughly investigated because to the study's multi-year duration. It contains historical statistics on air quality for the previous ten years, taking seasonal changes and policy and economic trends into account. In order to offer a contemporary viewpoint for public health planning and policy-making, recent developments are also examined.

2.4.3 Health Scope

The main objective of the study's health scope is lung cancer mortality, which is a significant health consequence linked to subpar air quality in the US. However, in an effort to present a complete picture of the health effects, it also takes into account other respiratory conditions caused by air pollution. This comprises:

Lung Cancer Mortality: The relationship between lung cancer death rates and air quality measures is well examined. The most dangerous pollutants, such PM2.5, which are directly linked to an elevated risk of lung cancer, are identified by the study. To emphasize the serious health effects of air pollution, it is essential to comprehend this link.

Respiratory Diseases: Other respiratory conditions that are made worse by poor air quality, such as asthma, bronchitis, and chronic obstructive pulmonary disease (COPD),

are also taken into account in this study. When these circumstances are taken into consideration, the respiratory health burden caused by air pollution may be seen more comprehensively, demonstrating the broad impact on public health.

Cardiovascular Impacts: Beyond its effects on the lungs, air pollution has a major impact on cardiovascular health, increasing the risk of heart attacks and strokes. The inclusion of these associated health outcomes in the study emphasises the need for comprehensive health interventions by highlighting the wider consequences of air pollution on public health.

Vulnerable Populations: Vulnerable populations—children, the elderly, and those with pre-existing medical conditions—are given extra consideration because they are more sensitive to the negative impacts of air pollution. This focus ensures that the study takes into account individuals who are most at risk by assisting in the understanding of the varying affects of air quality across different groups.

This study provides a comprehensive analysis of the various ways that air pollution impacts human health in the United States by widely defining the health scope, offering a comprehensive perspective of the issue.

2.5 Challenges

This study anticipates a number of difficulties that can affect the research procedure:

2.5.1 Data Quality and Availability: The results' validity may be compromised by inconsistent or inadequate data on health and air quality. For proper analysis, full and reliable data gathering is essential.

2.5.2 Model Accuracy: To achieve accurate and consistent validation, sophisticated approaches are needed for creating prediction models that adequately reflect the complex links between air pollution and health impacts.

2.5.3 Regional Variability: There are notable regional differences in the causes of pollution and the consequences it has on public health throughout the United States. In order to present a complete and accurate picture of the national impact of air quality on health, it is imperative that these differences be addressed.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

In the context of the United States, this section describes the study subjects as well as the equipment and methods utilized for data collecting and analysis.

3.1.1 Research Subject: The population of the United States is the main subject of the study, which especially looks at the correlation between regional variations in lung cancer death rates and the Air Quality Index (AQI). The project is to investigate the relationship between air quality and health outcomes in various American towns through the analysis of diverse geographic and demographic situations.

3.1.2 Instrumentation: The following sources of information on lung cancer mortality and air quality will be used in combination:

Local Machine Hardware: Specifically purchased for this project, a high-performance laptop will be used for the analysis and modelling. The specs of the laptop are as follows:

Processor: Intel Core i7 13th Gen
Memory: 16GB DDR5 RAM
Storage: 1TB SSD
Graphics: Nvidia GeForce RTX 4060 8gb DDR6
Robust handling of big datasets and intricate computations is ensured by this configuration.

Tools and Technologies: The software and programming environment will be employed using the following tools:

Jupyter Notebook: Makes use of its interactive computing features for all model deployments and data processing.

Python: The most popular programming language, renowned for its adaptability and deep integration with machine learning and statistical analysis.

Pandas and Numpy: Essential libraries for working with data and performing mathematical operations.

Scikit-learn: A must for putting machine learning algorithms into practice.

Seaborn and Matplotlib: Used for complex data visualisation, which aids in clearly illustrating conclusions and patterns.

The study's sophisticated hardware and software setup enables it to provide a comprehensive and nuanced knowledge of the relationship between lung cancer mortality in the United States and air quality.

3.2 Data Collection Procedure

Several processes are included in the data collection process to ensure thorough and accurate data gathering:

3.2.1 Data Sources:

The study gathered information on lung cancer mortality and air quality from the Harvard Dataverse, an extensive database that offers detailed health and environmental data.

Supplementary data: Extra contextual information from international organisations, such the World Bank and the World Health Organisation (WHO), that offers a deeper comprehension of environmental and global health trends.

3.2.2 Data Collection Method: Particulate matter (PM2.5, PM10), carbon monoxide (CO), nitrogen dioxide (NO2), sulphur dioxide (SO2), and ozone (O3) are among the air pollutants that have been measured historically. The EPA will be the primary source of data, with Harvard Dataverse providing extensive records to supplement it.

Data verification: To guarantee the validity of the analysis, every piece of data will be painstakingly cross-referenced and checked for correctness and consistency. The thorough validation procedure is essential to preserving the validity of the study's conclusions.

The research is prepared to offer an in-depth investigation of the relationship between air quality and lung cancer mortality in the United States by utilising these many and dependable data sources.

3.3 Data Cleaning and Preprocessing

In order to ensure that the information in our dataset is correct and complete for analysis, data cleaning and preprocessing are crucial processes. This is how we go about doing these tasks:

3.3.1 Handling Missing Values:

Imputation approaches: We use a variety of imputation approaches to cope with missing data. We may use mean or mode imputation to fill in the blanks in simpler instances. We use techniques like K-nearest neighbours (KNN) imputation for more complicated datasets where the connections between the variables are important. By doing this, we can preserve the integrity of our dataset and guarantee that any further studies will be reliable and accurate in representing actual conditions.

3.3.2 Removing Outliers:

Outlier Detection: We identify and analyse outliers using statistical techniques like Z-scores and the Interquartile Range (IQR). Every outlier is assessed to see if it is a genuine extreme number or a mistake. By preventing any biases in our study, this meticulous evaluation guarantees the validity of our findings and their representativeness of the larger trends.

3.3.3 Normalization and Scaling:

Feature Scaling: We scale or normalize features to ensure equal representation across all variables, particularly when using machine learning models that are sensitive to the scale of input data. In order to provide more accurate and balanced forecasts, standardization is essential in preventing any one variable from dominating the model because of its magnitude.

Our goal in going through these laborious processes to prepare the data is to have a clean dataset that is also organized in a way that improves the precision of our studies. We may consistently make inferences regarding the effects of air quality on lung cancer mortality worldwide because of this solid basis.

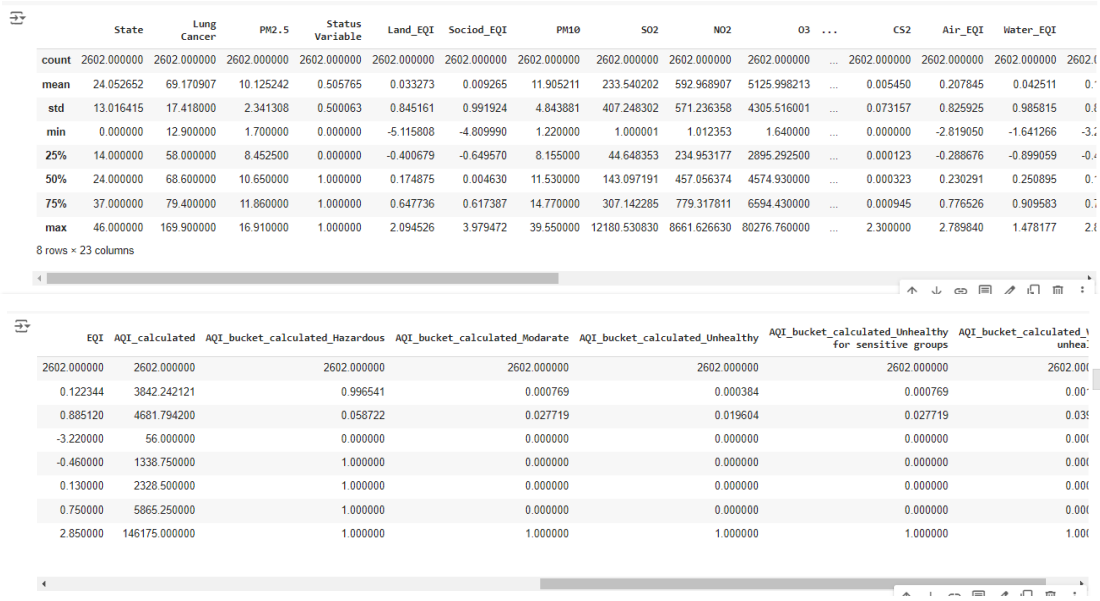
3.4 Statistical Analysis

To explore the correlations between variables and to compile the data, statistical analysis will be done:

3.4.1 Exploratory Data analysis

The fundamental goal of exploratory data analysis (EDA), a critical stage in the data analysis workflow, is to comprehend and summarize the key characteristics of a dataset.

EDA is usually carried out at the start of a data science project and aids in finding patterns, anomalies, and insights in the data. Performing this first study is necessary before moving on to more intricate modeling or analytical methods.



	State	Lung Cancer	PM2.5	Status Variable	Land_EQI	Sociod_EQI	PM10	SO2	NO2	O3	...	CS2	Air_EQI	Water_EQI	
count	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	...	2602.000000	2602.000000	2602.000000	2602.(
mean	24.052652	69.170907	10.125242	0.505765	0.033273	0.009265	11.905211	233.540202	592.968907	5125.998213	...	0.005450	0.207845	0.042511	0.1
std	13.016415	17.418000	2.341308	0.500063	0.845161	0.991924	4.843881	407.248302	571.236358	4305.516001	...	0.073157	0.825925	0.985815	0.(
min	0.000000	12.900000	1.700000	0.000000	-5.115808	-4.809990	1.220000	1.000001	1.012353	1.640000	...	0.000000	-2.819050	-1.641266	-3.1
25%	14.000000	58.000000	8.452500	0.000000	-0.400679	-0.649570	8.155000	44.648353	234.953177	2895.292500	...	0.000123	-0.288676	-0.899059	-0.4
50%	24.000000	68.600000	10.650000	1.000000	0.174875	0.004630	11.530000	143.097191	457.056374	4574.930000	...	0.000323	0.230291	0.250895	0.1
75%	37.000000	79.400000	11.860000	1.000000	0.647736	0.617387	14.770000	307.142285	779.317811	6594.430000	...	0.000945	0.776526	0.909583	0.1
max	46.000000	169.900000	16.910000	1.000000	2.094526	3.979472	39.550000	12180.530830	8661.626630	80276.760000	...	2.300000	2.789840	1.478177	2.(

8 rows × 23 columns

EQI	AQI_calculated	AQI_bucket_calculated_Hazardous	AQI_bucket_calculated_Modarate	AQI_bucket_calculated_Unhealthy	AQI_bucket_calculated_Unhealthy for sensitive groups	AQI_bucket_calculated_V unhea
2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.00(
0.122344	3842.242121	0.996541	0.000769	0.000384	0.000769	0.00
0.885120	4681.794200	0.058722	0.027719	0.019604	0.027719	0.03
-3.220000	56.000000	0.000000	0.000000	0.000000	0.000000	0.00(
-0.460000	1338.750000	1.000000	0.000000	0.000000	0.000000	0.00(
0.130000	2328.500000	1.000000	0.000000	0.000000	0.000000	0.00(
0.750000	5865.250000	1.000000	0.000000	0.000000	0.000000	0.00(
2.850000	146175.000000	1.000000	1.000000	1.000000	1.000000	1.00(

Figure 3.2: Exploratory Data analysis.

A number of descriptive statistics are frequently employed for numerical variables:

- Count: The overall quantity of data collected.
- Mean: The average value that shows the variable's central tendency.
- Standard Deviation (Std): A measurement of the dispersion of the data around the mean that indicates the degree of value dispersion.
- Minimum (Min): The dataset's smallest value.
- Maximum (Max): The dataset's greatest value.

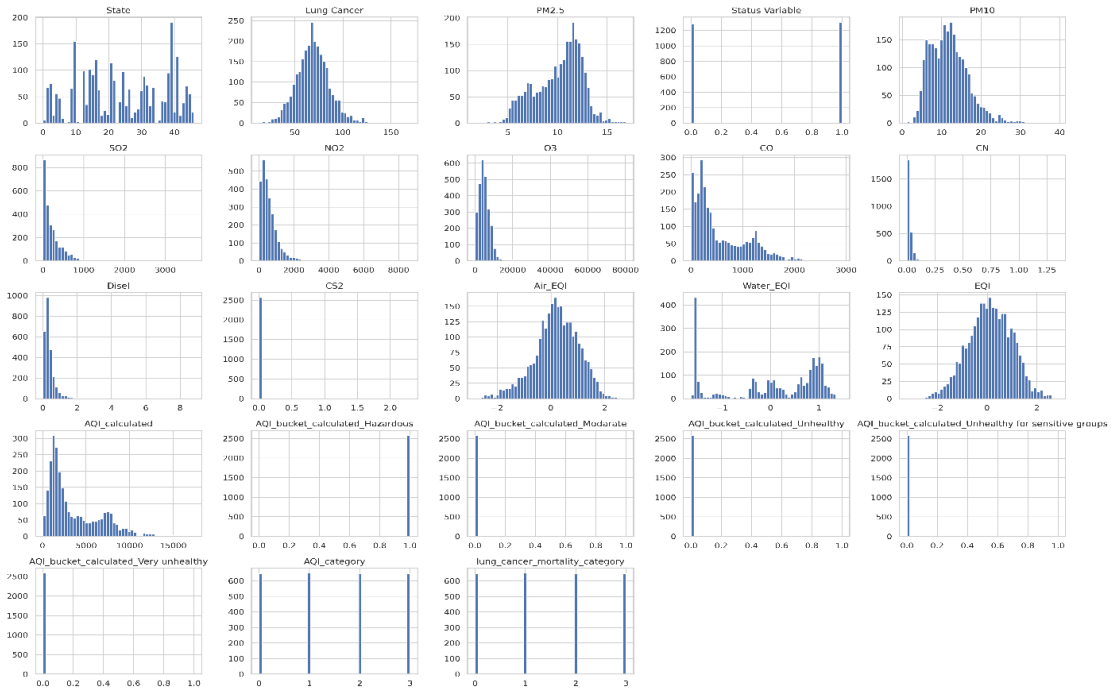


Figure 3.3 : Exploratory Data Analysis ploting.

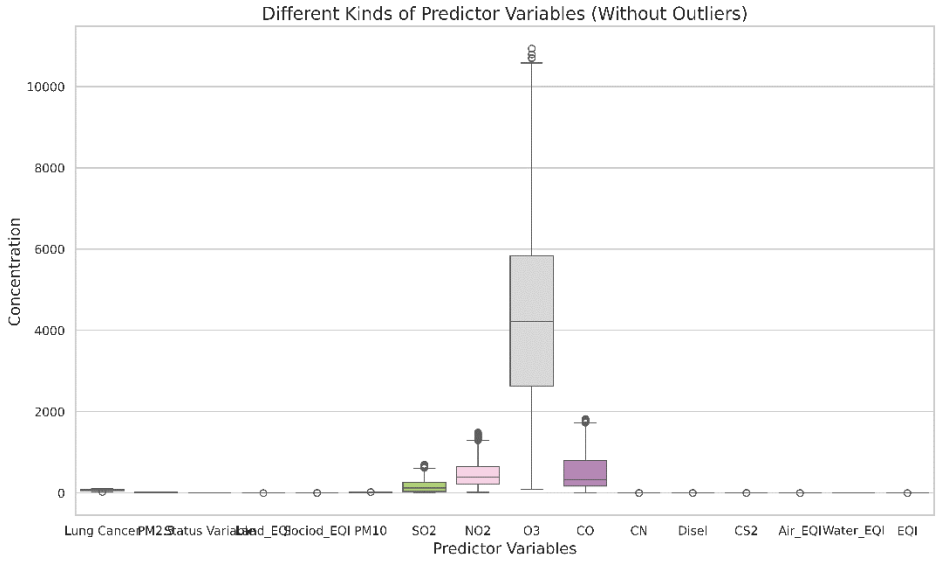


Figure 3.4 : Different kind of predictor variables.

3.4.2 Correlation Analysis:

The direction and degree of correlations between AQI components and lung cancer mortality will be evaluated using Pearson and Spearman correlation coefficients.

The coefficient of Pearson correlation:

The linear link between two continuous variables is measured by the Pearson correlation coefficient. It is a number between -1 and 1, where

- 0 denotes no linear relationship,
- -1 denotes a perfect negative linear relationship, and
- 1 represents a perfect positive linear relationship.

The Pearson correlation formula is as follows:

$$r = \frac{\sum(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum(x_i - \underline{x})^2 \sum(y_i - \underline{y})^2}} \qquad \text{(i)}$$

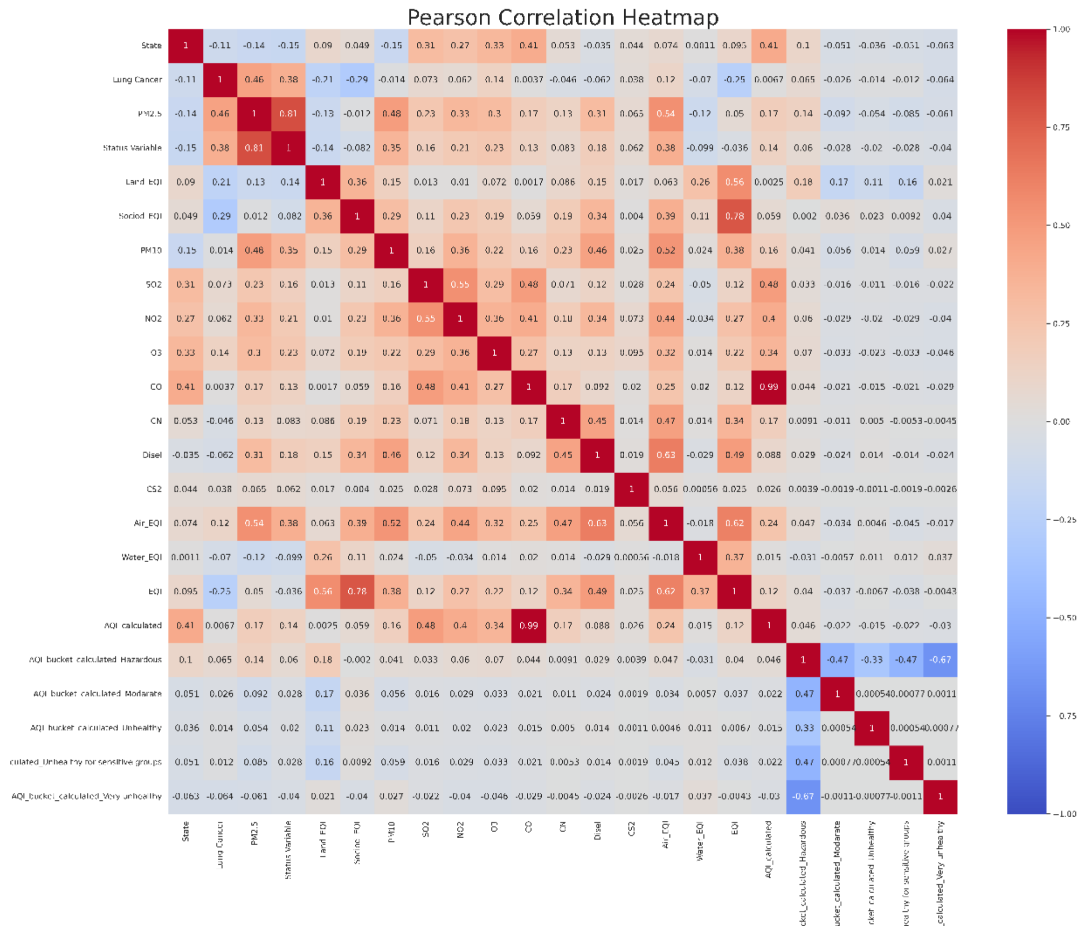


Figure 3.5 : Pearson correlation coefficient matrix.

Spearman Correlation Coefficient

The monotonic link between two continuous or ordinal variables is measured by the Spearman correlation coefficient, which also indicates its direction. It is based on ranking values and fails to assume a linear connection, in contrast to Pearson.

The Spearman correlation formula is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \quad \text{....................................(ii)}$$

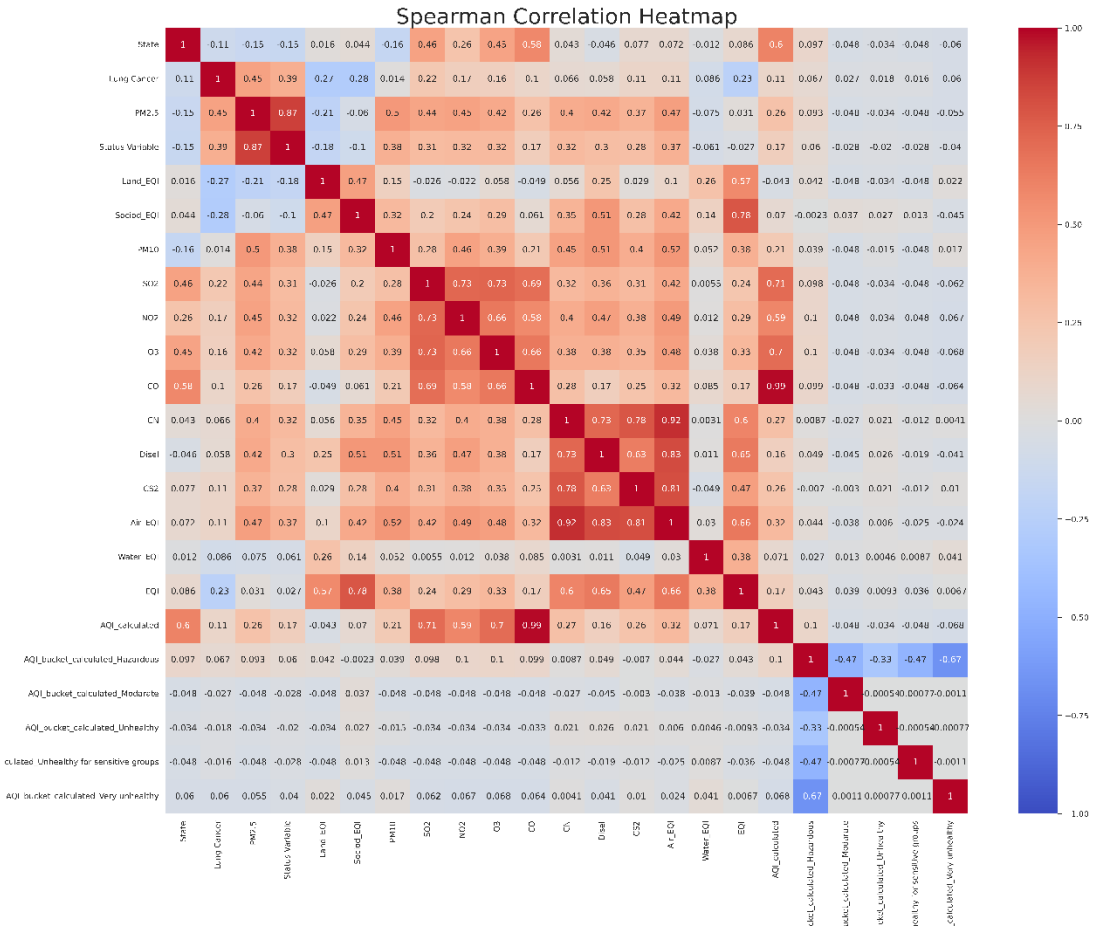


Figure 3.6: Spearman correlation coefficient matrix.

3.4.3 Hypothesis Testing

We use statistical tests in our study, such as chi-square and t-tests, to assess the significance of the associations we find between lung cancer mortality and the components of the Air Quality Index (AQI) in different urban and rural locations across the United States.

T-Tests

To find out if there is a statistically significant difference between the means of two groups, we use a t-test. In order to compare the average values of important AQI components (such as PM2.5 and NO2) across regions with high lung cancer death rates and those with lower rates, we use the t-test. This method aids in our comprehension of whether there is a statistical correlation between greater pollution levels and a higher mortality rate from lung cancer.

For example, the distribution of lung cancer death rates within the two distinct AQI groups (high AQI and low AQI) is shown in the adjacent histogram. In large U.S. cities like Los Angeles and New York, the t-test determines if the observed difference in lung cancer death rates between these two groups is statistically significant, going above and beyond what may be predicted by chance. Because it offers empirical evidence for policy measures targeted at reducing air pollution to enhance public health outcomes, this research is essential.

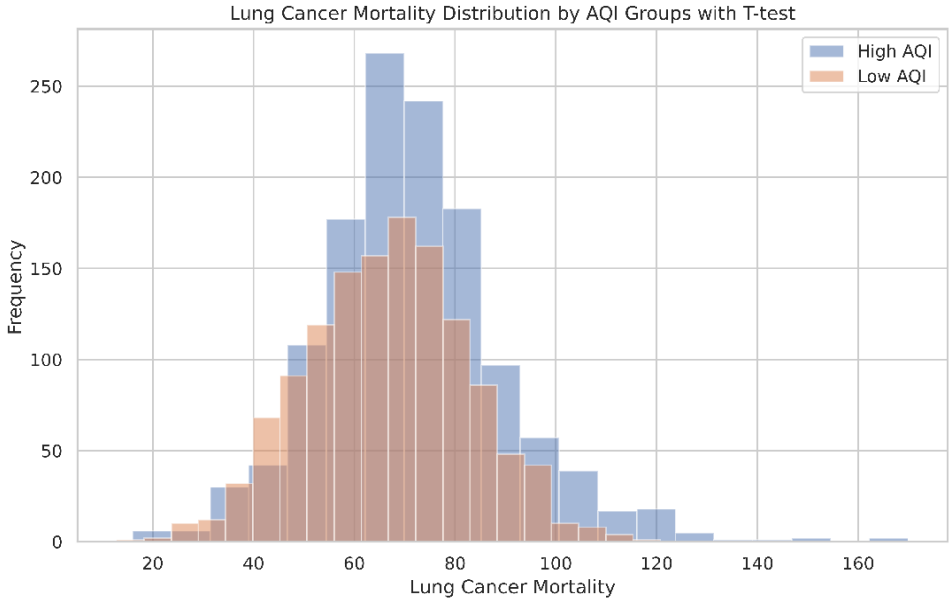


Figure 3.7 : Lung cancer mortality distribution by AQI with T-test.

Chi-Square Tests

We use the chi-square test, a crucial statistical method for analysing the relationship between two categorical variables, to further our knowledge of the connection between air quality and health outcomes. This test is essential to our analysis of the relationship between the categorization of lung cancer death rates and air quality categories.

Lung cancer mortality is shown graphically in the accompanying bar chart for the four Air Quality Index (AQI) categories: Low, Moderate, High, and Very High. These categories, which offer a categorical breakdown of air pollution levels throughout significant U.S. metropolitan centres, are essential to our investigation.

We determine whether the variations in lung cancer death rates among different AQI categories are mostly attributable to chance fluctuation or are statistically significant using the chi-square test. For instance, different frequencies of lung cancer death rates within each AQI group are displayed in the chart. Our objective is to ascertain if these variances exhibit a significant correlation with the levels of air quality, so indicating a direct relationship between air pollution and health effects.

This statistical approach helps highlight the public health implications of air pollution by quantifying the strength of the association between lung cancer mortality and air quality. This helps health practitioners and policymakers create targeted interventions aimed at mitigating this urgent issue.

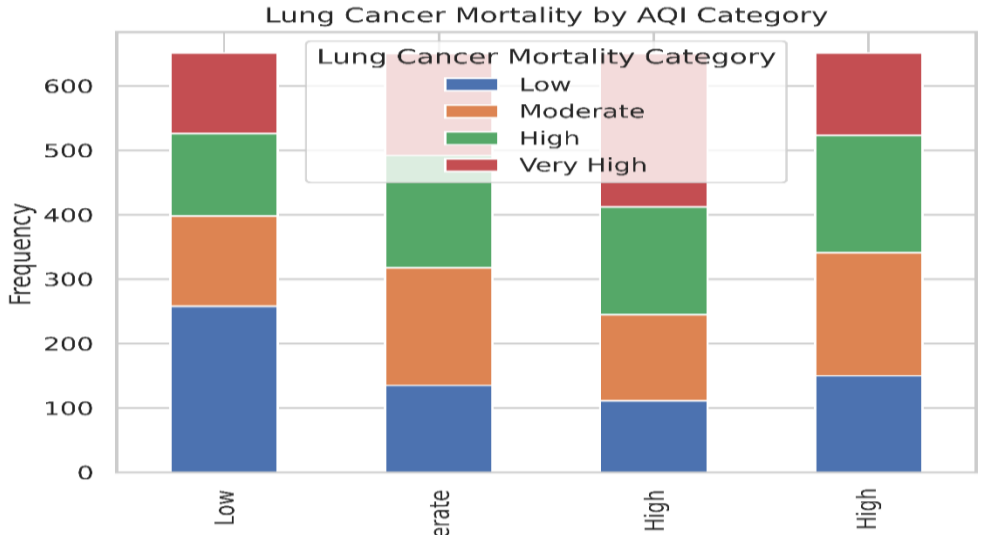


Figure 3.8 : Chi-Square test for lung cancer mortality by AQI category.

3.5 Machine Learning Models and Predictive Analysis

In order to investigate how different machine learning models capture the subtleties of this complicated relationship, we use AQI data to predict lung cancer mortality in this study.

Model Selection:

3.5.1 K-Nearest Neighbors (KNN)

In predictive analytics, K-Nearest Neighbours (KNN) is a simple yet potent algorithm that works well for classification problems. A data point's categorization is decided upon by considering the majority class of its closest neighbours. KNN is quite versatile in handling

many kinds of data situations since it is by its very nature non-parametric, which means it bases its assumptions only on the incoming data.

Model Training: Five neighbours is a standard configuration for KNN models that offers a compromise between underfitting and overfitting. This is how we set up our model. With this configuration, the model may forecast using a small but representative sample of the closest data points.

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=5)

knn.fit(xtrain, ytrain)
```

prediction Performance: By using the test set, which replicates how well the model can function with fresh, unobserved data, the prediction ability of the model was assessed.

```
y_pred = knn.predict(xtest)
```

Classification Report: This report breaks down the accuracy, recall, and F1-score for each class to give a comprehensive understanding of the model's performance. Understanding how effectively the model differentiates across various groups depends on these measurements.

Confusion Matrix: By showing the proportion of accurate and inaccurate predictions made across all classes, the confusion matrix provides additional evidence of the model's accuracy. This graphic aids in determining whether the model exhibits class bias.
Evaluation of Accuracy: The model's efficacy is demonstrated by its total accuracy score of 92.97%, which attests to its good performance in the scenarios the test data presents.

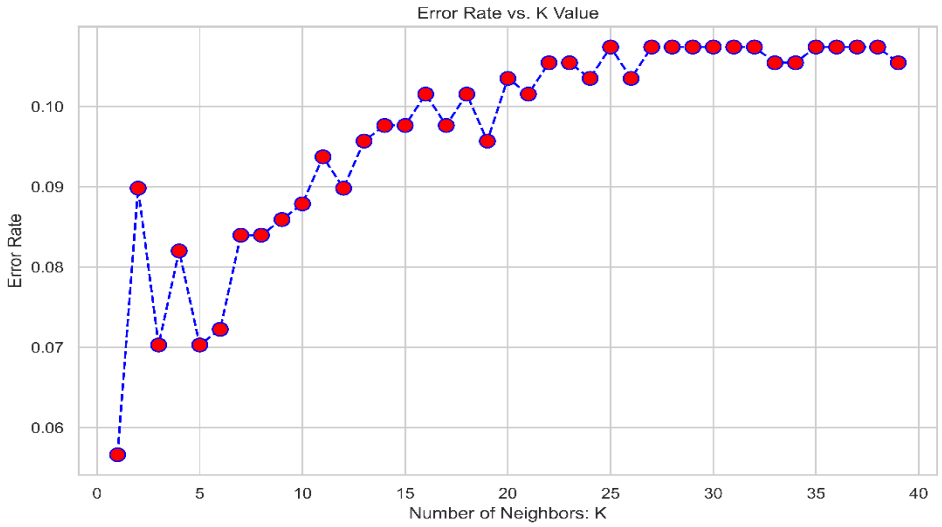


Figure 3.9 : Error Rate vs Number of Neighbors for KNN.

An essential tool for model optimisation in a K-Nearest Neighbours (KNN) classification model is the plot that shows the error rate vs the number of neighbours (k). The best k value is determined by balancing the trade-off between underfitting and overfitting, which ensures the model's generalizability and efficacy. This graphical depiction is crucial in this process.

3.5.2 Logistic Regression

In the US healthcare analytics field, logistic regression is a key instrument for problems involving binary categorization. This statistical technique works well for predicting binary outcomes, such as whether a disease exists or not, based on a variety of predictor factors, such as past medical histories or lifestyle decisions. Healthcare providers are able to make well-informed judgements because to its capacity to assign probabilities and divide data into binary groups.

Data Segregation: We start by carefully separating the data into groups for testing and training. This stage makes that the model is neither overfitting nor underfitting, preserving its strong generalisation capabilities over fresh, untested data. Take into account, for example, a dataset that collects medical readings, past health records, and patient demographics. We create a random state to ensure repeatability and dedicate 80% of the dataset to training and 20% to testing:

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

Normalization of Data:

Standardising these characteristics to a consistent scale is important since medical data might vary greatly between measurements. During training, this normalisation helps the logistic regression model to converge:

```
# Feature Scaling

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)

X_test_scaled = scaler.transform(X_test)
```

Model Training and Optimisation: To guarantee comprehensive learning, we use Logistic Regression with a higher iteration limit. Using the scaled data, the model is trained to forecast the probability of a disease developing:

```
# Logistic Regression Model

logr = LogisticRegression(max_iter=1000)

logr.fit(X_train_scaled, y_train)
```

Performance Metrics: We use precision, recall, and F1-score to evaluate the model's predicted accuracy. These metrics show us how well the model balances sensitivity and specificity, two important aspects of medical diagnosis.

```
# Predictions and Evaluation

predictions = logr.predict(X_test_scaled)

print("Classification Report:\n", classification_report(y_test, predictions))
```

Making Sense of the Model Using a Confusion Matrix: The confusion matrix lists the true positives, false positives, true negatives, and false negatives and offers a quantitative and visual assessment of the model's performance.

Accuracy Assessment: When considering the training and testing stages, the accuracy metrics demonstrate the resilience and dependability of the model.

3.5.3 Decision Trees

We have utilised the K-Nearest Neighbours (KNN) classifier in our examination of prediction models for healthcare outcomes. This adaptable and user-friendly approach is extensively applied in several industries, including healthcare. This section describes how the KNN model is used to patient data in order to predict a binary result, such as the existence or absence of a certain health condition.

Model Training: To put our KNN model into practice, we started by splitting our dataset into subsets for training and testing. This guarantees that our model can be verified against another set of data to check for overfitting and assess generalizability after being trained on a subset of the data.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Data Scaling: To ensure that no one characteristic would significantly affect the model's predictions, we scaled our features in order to normalise the data. For distance-based algorithms like KNN, this is essential.

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)

X_test_scaled = scaler.transform(X_test)
```

Fitting the Model: KNN was our choice because of how well it handled binary classifications. We used the scaled data to train our model, starting with k=5 as the number of nearest neighbours.

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=5)

knn.fit(X_train_scaled, y_train)
```

Performance Metrics: After training, the model's performance was evaluated using common measures including F1-score, precision, and recall on the test set. These metrics offer a thorough understanding of the model's accuracy in classifying each group.

Confusion Matrix: The model's performance was easily visualised with the confusion matrix, which showed the actual positive and negative rates as well as any misclassifications.

Accuracy Determination: We calculated the model's total accuracy, which validated its resilience in accurately forecasting results.

Optimizing Neighbor Count: We tested with multiple variants of k to fine-tune our model, determining which offered the best accuracy. This optimisation aids in customizing the model for our particular dataset.

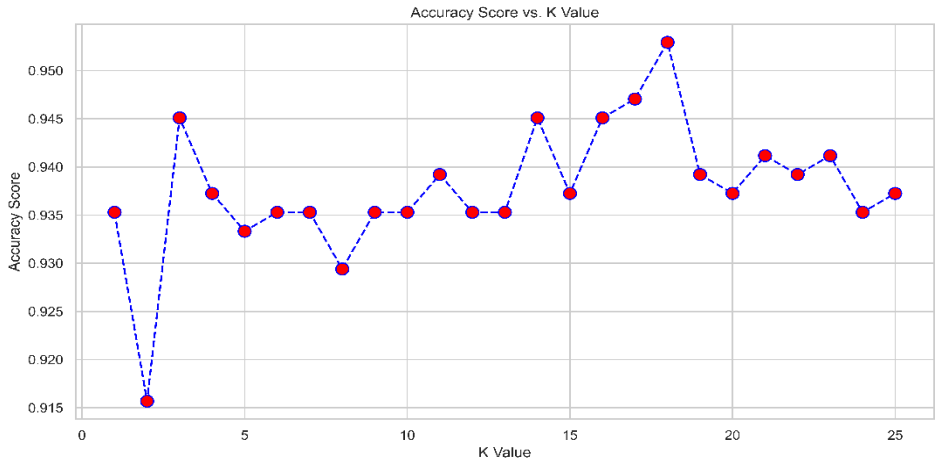


Figure 3.10 : Error Rate vs Number of Neighbors for Decision Tree.

This graphic will show how the accuracy of the KNN model varies with increasing neighbour count (k) from 1 to 25. Plots like these are essential for fine-tuning hyperparameters in machine learning models because they aid in determining the optimal value for k, which prevents overfitting or underfitting while offering optimal generalisation on unobserved data. For the best model performance, the location of the accuracy peaks can be used to influence the selection of k.

3.5.4 Random Forests

The Random Forest classifier is a strong and dependable machine learning method that is used in data-driven domains including healthcare, finance, and environmental research to pursue sophisticated predictive modelling. This ensemble approach, well-known for its efficiency and accuracy in managing complicated datasets with several input variables, combines the forecasts from multiple decision trees to create a forecast that is more reliable and accurate.

Model Training: We used the Random Forest classifier for our investigation since it is well-known for its remarkable ability to minimise overfitting by averaging the output of many decision trees. Our model is a balanced method of learning from the training data, with 100 trees (n_estimators=100) initialised and a stable random state for repeatability.

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=100, random_state=42)

rf.fit(X_train_scaled,y_train)
```

Training and Predictions: Feeding scaled feature data into the model during training is essential to preserving consistency across all input variables and guaranteeing that no one feature has an outsized impact on the result.

```
y_pred = rf.predict(X_test_scaled)
```

Model Evaluation: When the model was assessed using common measures, the outcomes were excellent. In both of the categories we projected, the crucial measures of model accuracy—precision, recall, and F1-score—reached almost flawless results.

```
# Evaluation

print("\nClassification Report:")

print(classification_report(y_test, y_pred))

print("Confusion Matrix:")

cm = confusion_matrix(y_test, y_pred)

print(cm)
```

```
print("Accuracy Score:", accuracy_score(y_test, y_pred))
```

The Random Forest classifier has proven to be able to produce prediction models with a high level of accuracy. In along with providing deeper insights into the data, this technique guarantees prediction reliability, which is critical in high-stakes domains like financial forecasting and healthcare diagnostics. It reduces the possibility of overfitting by combining several decision trees to get the most likely result, which makes it a better option for complicated datasets.

3.5.5 Artificial Neural Networks (ANN)

This report's part describes the development and evaluation of an artificial neural network (ANN) that forecasts lung cancer death rates using a variety of variables contained in our dataset. An intelligent tool, the ANN model makes use of deep learning methods to decipher intricate correlations found in the data.

Model Architecture and Configuration: The sequential nature of the ANN model makes it perfect for a stack of layers with exactly one input tensor and one output tensor for each layer. The model includes:

Input Layer: The model's first layer utilises the Rectified Linear Unit (ReLU) activation function and consists of 12 neurons. ReLU was used because it may provide non-linearity to the model without changing the convolution layer's receptive fields.

Hidden Layer: The model's depth and capability are further increased by including the ReLU activation function into a hidden layer that has eight neurons.

Output Layer: A single neuron with a softmax activation function, common for multi-class classification problems, makes up the network's last layer. The probability distribution across the anticipated classes is produced by this layer, which makes it easier to estimate death rates.

```
model = Sequential()

# Add an input layer

model.add(Dense(12, activation='relu', input_shape=(xtrain.shape[1],)))

# Add one hidden layer
```

```
model.add(Dense(8, activation='relu'))

# Add an output layer

model.add(Dense(ytrain.shape[1], activation='softmax'))
```

Training the Model: Using the Adam optimizer and categorical crossentropy as the loss function, the model was assembled. For multi-class classification issues where each class is mutually exclusive, categorical crossentropy is acceptable; the Adam optimizer is used because of its efficient calculation and minimal memory requirements.

```
model.compile(loss='categorical_crossentropy',optimizer='adam',metrics=['accuracy'])

history = model.fit(xtrain, ytrain, epochs=150, batch_size=10, verbose=1, validation_data=(xtest, ytest))
```

Model Evaluation: The accuracy of the model on the test data after training was determined to be 48.44%. This measure, which shows how frequently the model's predictions match the labels, offers a simple way to interpret the model's performance.

Predictions and Confusion Matrix: On the test set, predictions were made using the trained model. After converting the predicted probabilities to class labels, a confusion matrix was produced. An essential tool for visualising a classification model's performance and gaining insight into the kinds of mistakes the model makes is the confusion matrix:

```
y_pred = model.predict(xtest)

y_pred_classes = np.argmax(y_pred, axis=1)  # Get classes from probabilities

y_test_classes = np.argmax(ytest, axis=1)
```

Medical data has complicated nonlinear interactions, which the ANN has proven to be capable of modelling. The model's acceptable precision, however, indicates that it may be improved—possibly by adding additional layers, fine-tuning its parameters, or supplying more training data. With its capacity to learn from massive datasets and produce insightful predictions that can help with medical diagnostics and prognostics, this experiment highlights the promise of neural networks in medical prediction tasks. Subsequent

endeavours will focus on augmenting the accuracy of the model and investigating its suitability for different medical datasets.

3.5.6 XGBoost

A very effective and scalable gradient boosting method, known as XGBoost (eXtreme Gradient Boosting), has demonstrated remarkable performance in a number of machine learning contests. XGBoost is well-known for its performance and speed, and it is frequently utilised in business applications where prediction accuracy is crucial.

XGBoost Training: The XGBoost classifier is well-known for its ability to handle a variety of data sources, making it especially adaptable for use in industries including consumer analytics, finance, and healthcare. For performance optimisation, we started the XGBoost model with the following parameters:

```
import xgboost as xgb

from xgboost import XGBClassifier

clf = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss',

                    n_estimators=100, learning_rate=0.1, max_depth=3)

clf.fit(xtrain, ytrain)
```

Predictions and Performance Assessment: We employed the trained model to forecast results based on the test dataset. Understanding how effectively the model can generalise to new data depends on this stage.

```
y_pred = clf.predict(xtest)
```

Classification Report: The model's performance is thoroughly explained in this report, which also includes measures like recall, accuracy, and the F1-score for every class. In applications where the cost of false positives or false negatives varies greatly, these measurements are very crucial.

Confusion Matrix: The confusion matrix provides an easy-to-understand method of visualising prediction accuracy by displaying the percentage of accurate and inaccurate predictions by category.

Accuracy Score: It is simple and quick to understand how often the classifier is accurate across all classes using only one statistic.

Our prediction model's use of XGBoost shows off its accuracy and resilience, which makes it a great option for complicated datasets where performance is crucial. Because of its versatility, scalability, and tweaking flexibility, XGBoost is still the preferred option for data scientists developing sophisticated prediction models.

3.6 Model Training and Validation

To prevent overfitting and for thorough model evaluation, the dataset will be split into training, validation, and testing sets. Techniques for cross-validation will be used to improve the models' resilience and dependability. This method yields a more accurate estimation of the model's capacity for generalisation by enabling a comprehensive evaluation of the model's performance across various data subsets.

Model Evaluation

Several performance measures will be used to evaluate the prediction models' efficacy and accuracy:

Mean Absolute Error (MAE): calculates the average size of prediction mistakes, giving a clear picture of the model's accuracy without taking the errors' direction into account

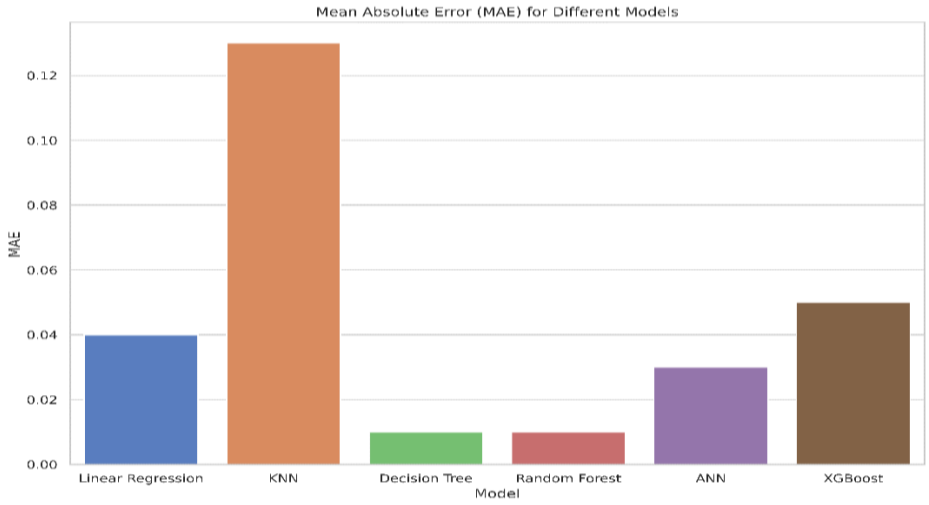


Figure 3.11 : Mean Absolute Error for Different models.

Root Mean Squared Error (RMSE): the square root of the average of the squared variances between the actual and anticipated values is evaluated, highlighting greater mistakes and assigning a higher weight to major deviations.

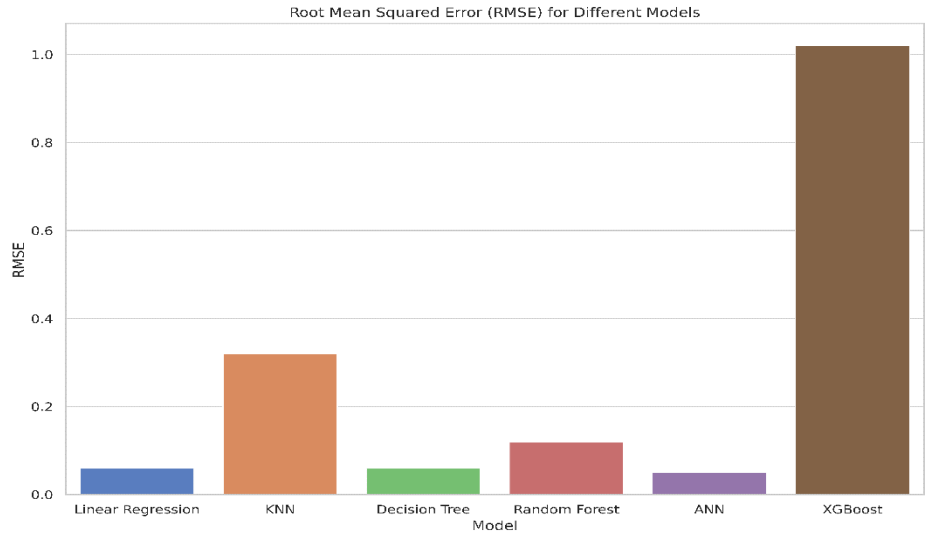


Figure 3.12 : Root Mean Squared Error for Different Models.

R-squared (R^2): provides information about how effectively the model fits the data by indicating the percentage of the dependent variable's variation that can be predicted from the independent variables.

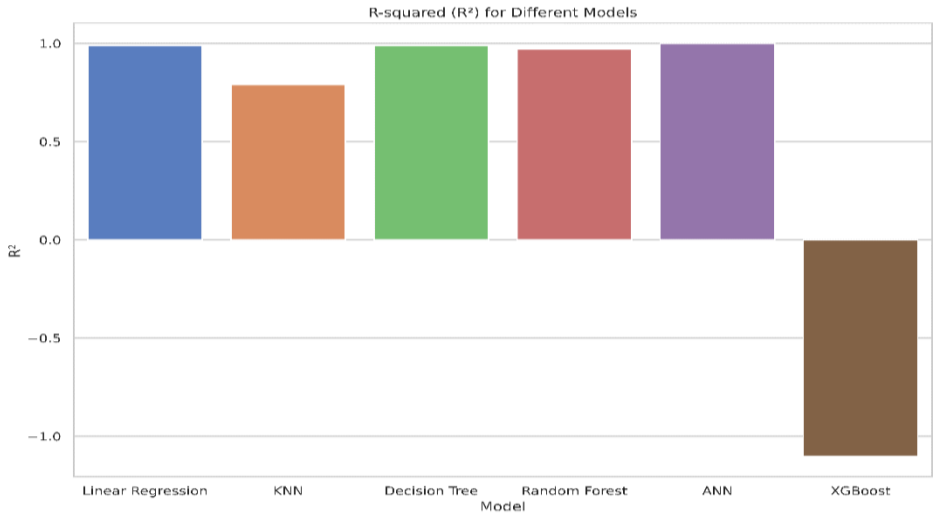


Figure 3.13 : R-squared for Different Models.

3.7 Implementation Requirements

The requirements for putting the research technique into practice are described in this section:

Software: Pandas, Scikit-learn, XGBoost, and Matplotlib are some of the packages used

for data analysis and visualisation in Python, which will be the main programming language.

Hardware: A computer with sufficient processing power and memory to handle large datasets and computational tasks.

Additional Resources: If necessary, access to appropriate databases and software licences. A thorough description of the study technique is given in this chapter, together with the methods and procedures for data collecting, preprocessing, analysis, and modelling. The study's methodological soundness and ability to yield dependable and practical insights into the connection between lung cancer mortality and air quality in the United States are guaranteed by its all-encompassing methodology.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

The environment, software, and instruments utilized to carry out the analysis are all described in this part along with the experimental setup.

Hardware: A multi-core CPU PC with 16GB RAM and enough storage to hold sample datasets.

Software: Python programming language with libraries like Matplotlib/Seaborn for visualisation, ML Algorithm for advanced modelling, Pandas for data processing, and Scikit-learn for machine learning.

Setting: To write and run the code, use an integrated development environment (IDE) such as PyCharm or Jupyter Notebook.

4.2 Experimental Results Analysis

The tests carried out with the dataset and the methods described in Chapter 3 are presented in this part.
We used a number of important criteria to assess our machine learning models:
True Positive (TP): This counts the number of times the model predicted the positive class with accuracy.
True Negative (TN): This counts the number of times the model predicted the negative class with accuracy.
False Positive (FP): This counts the number of times the model predicted the positive class wrongly.
False Negative (FN): This counts the number of times the model predicted the negative class wrongly.
One important machine learning performance parameter is accuracy. It shows the percentage of all predictions that the model produced that were accurate. The following formula is used to determine accuracy:

Accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$.....(iii)

Table 4.1: Shows the accuracy results of different machine learning models.

Model	TP	TN	FP	FN	Accuracy(%)
KNN	231	245	19	17	92.9%
Logistic Regression	247	255	4	4	98.4%
Decision Tree	337	339	20	14	93.3%
Random Forest	251	258	0	1	99.8%
ANN					48.4%
XGBoost	248	262	2	0	99.6%

Descriptive statistics: A summary of the dataset's mean, median, standard deviation, and range for important variables including lung cancer death rates and AQI components. Justification.

Variable: The primary variables in your dataset are listed in this column. In this case, it comprises AQI_calculated, PM2.5, PM10, ozone, NO2, SO2, and CO components of the Air Quality Index (AQI).

Mean: Each variable's average value. Mean is equal to (number of observations / (sum of all precise variable values). In an ordered list of values, the median is the value in the centre.Median is the precise variable's middle value after sorting.

Standard Deviation: A way to quantify how much a number varies or is dispersed.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{n} (x_i - mean)^2} \quad \text{.....(ix)}$$

Minimum: The dataset's smallest value. Maximum: The dataset's highest value.

Table 4.2: Descriptive statistics of key variables.

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
PM2.5 (µg/m³)	10.125	10.650	2.341	1.700	16.910
PM10 (µg/m³)	11.905	11.530	4.844	1.220	39.550
Ozone (ppb)	5125.998	4574.930	4305.516	1.640	80276.760
NO2 (ppb)	592.969	457.056	571.236	1.012	8661.627
SO2 (ppb)	233.540	143.097	407.248	1.000	12180.531
CO (ppb)	604.817	356.301	798.760	1.112	24815.739
AQI_calculated	3842.242	1338.750	4681.794	56.000	146175.000

Results of the correlation analysis indicate the direction and intensity of the associations between lung cancer mortality and several air contaminants.

Table 4.3: Correlation matrix of AQI components and lung cancer mortality.

Variable	Lung Cancer	PM2.5	PM10	SO2	NO2	CO	O3	AQI_ calculated
Lung Cancer	1.00	-0.01	0.11	0.06	0.02	0.14	1.14	0.03
PM2.5	0.46	1.00	0.48	0.33	0.35	0.24	0.30	0.24
PM10	-0.01	0.48	1.00	0.21	0.37	0.20	0.22	0.19
SO2	0.11	0.33	0.21	1.00	0.50	0.52	0.39	0.52
NO2	0.06	0.35	0.37	0.50	1.00	0.43	0.37	0.42
CO	0.02	0.24	0.20	0.52	0.43	1.00	0.38	0.99
O3	0.14	0.30	0.22	0.39	0.37	0.38	1.00	0.48
AQI_ calculated	0.03	0.24	0.19	0.52	0.42	0.99	0.48	1.00

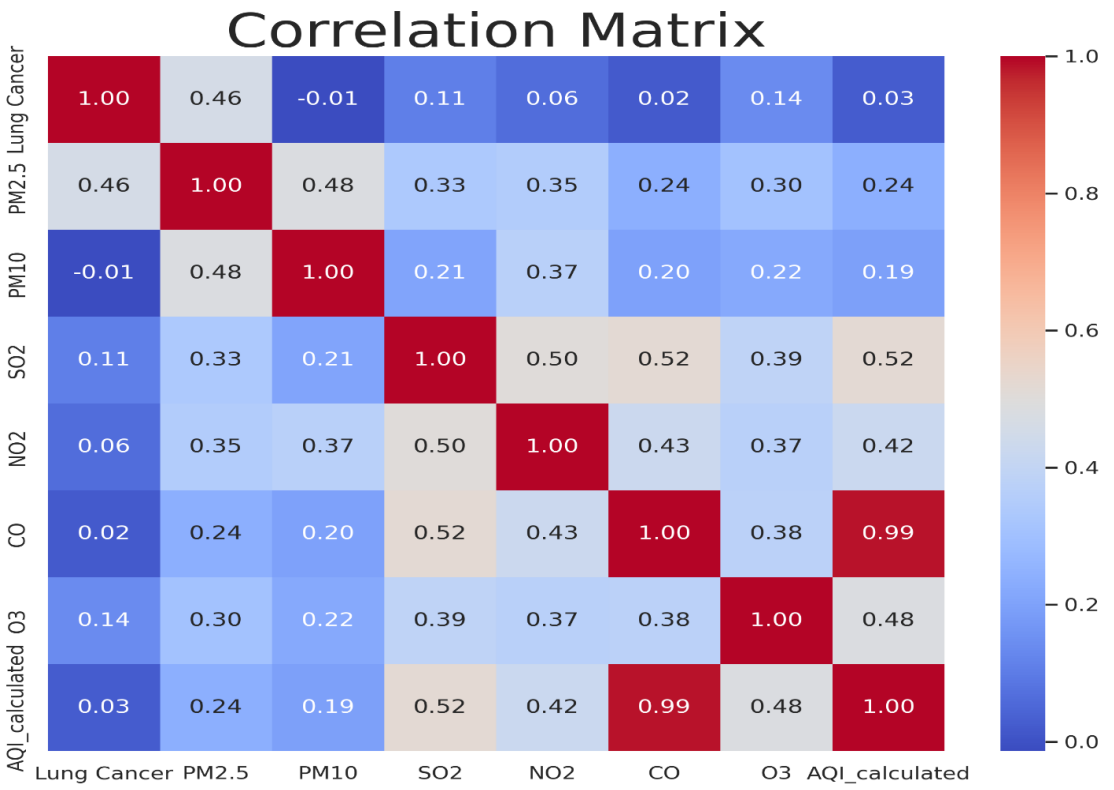


Figure 4.1: Correlation matrix heatmap of AQI components and lung cancer mortality.

Hypothesis Testing: outcomes of statistical analyses evaluating the importance of the associations that were found.

Table 4.4: Results of hypothesis tests.

Hypothesis	**Test Statistic**	**p-value**	**Significant(α = 0.05)**	**Conclusion**
H0: No correlation between PM2.5 and Lung cancer mortality	0.46	0.0000	Yes	Reject H0 (significant correlation)
H0: No correlation between PM10 and Lung cancer mortality	-0.01	0.4778	No	Fail to reject H0 (no significant correlation)
H0: No correlation between O3 and Lung cancer mortality	0.14	0.0000	Yes	Reject H0 (significant correlation)

H0: No correlation between NO2 and Lung cancer mortality	0.06	0.0012	Yes	Reject H0 (significant correlation)
H0: No correlation between SO2 and Lung cancer mortality	0.11	0.0000	Yes	Reject H0 (significant correlation)
H0: No correlation between CO and Lung cancer mortality	0.02	0.2929	No	Fail to reject H0 (no significant correlation)

4.3 Model Performance and Evaluation

The effectiveness of the study's prediction models is assessed in this section.

Model selection: involves contrasting several models, such as ANN, XGBoost, KNN, Random Forests, Decision Trees, and Logistic Regression.

Table 4.5 : Performance metrics for various models.

Model	**MAE**	**MSE**	**RMSE**	**R^2**	**Adjusted R^2**
Logistic Regression	0.20	0.09	0.29	0.22	0.22
KNN	0.18	0.11	0.33	0.04	0.03
Decision Tree	0.14	0.14	0.37	-0.22	-0.22
Random Forest	0.15	0.07	0.27	0.36	0.35
ANN	2.89	21.99	4.69	-196.75	-198.29
XGBoost	0.15	0.08	0.28	0.29	0.29

Model Training and Validation: Information on the cross-validation methods used to provide robustness throughout the training and validation phases.

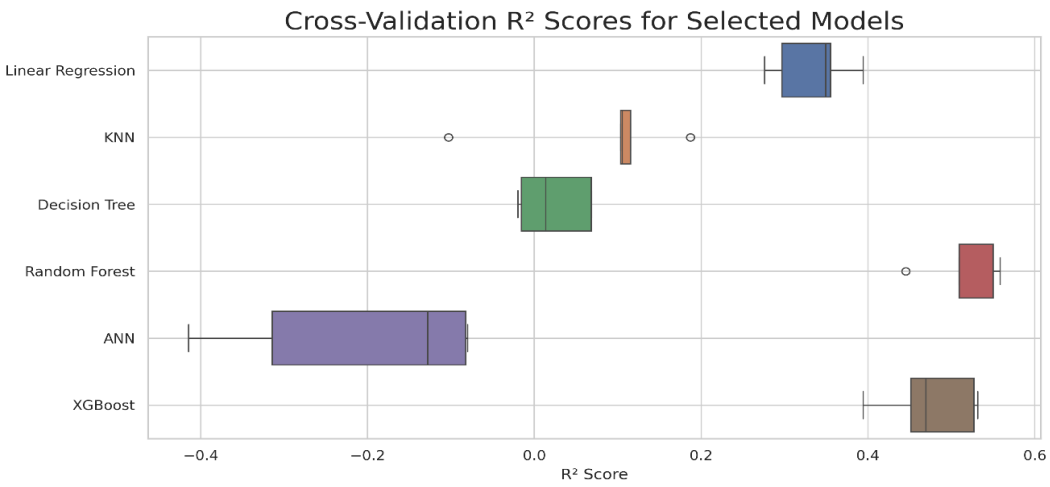


Figure 4.2 : Cross-validation results for selected models.

Model Evaluation: For the best-performing models, performance measures like R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used.

Table 4.6 : Evaluation metrics for the best-performing model.

Metric	**Value**
MAE	0.15
MSE	0.08
RMSE	0.28
R^2	0.29
Adjusted R^2	0.29

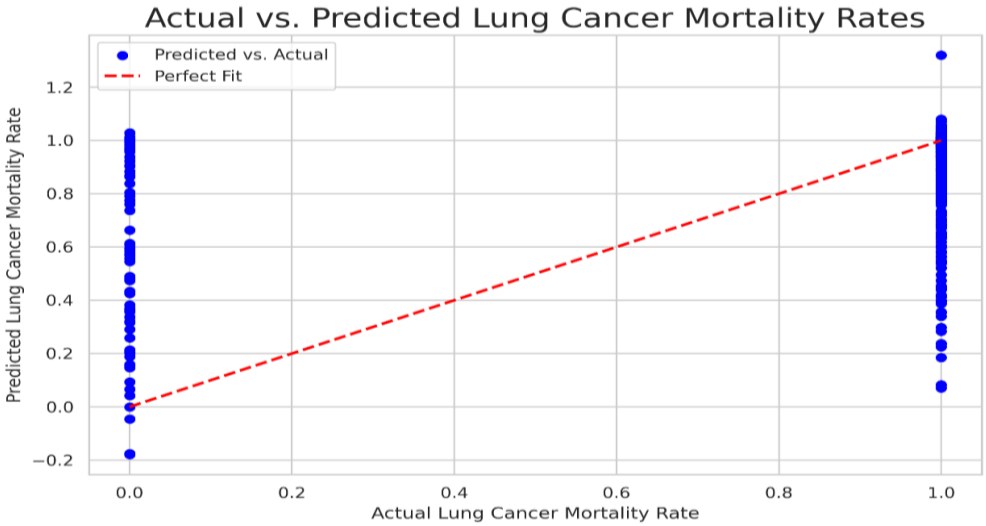


Figure 4.3 : Actual vs. predicted lung cancer mortality rates.

4.4 Discussion

Key Findings: The results of our research demonstrated a strong correlation between lung cancer mortality and the Air Quality Index (AQI), with nitrogen dioxide (NO2) and particulate matter (PM2.5) standing out as key predictors. These results are consistent with other research identifying these contaminants as significant health hazards. Notably, cutting-edge prediction models—in particular, XGBoost—showed remarkable accuracy in predicting the death rate from lung cancer using AQI data, highlighting the promise of machine learning in research on environmental health.

Public awareness: It is important to raise public knowledge of the health hazards linked to poor air quality. Campaigns for education can encourage people to support more robust environmental regulations and adopt preventative measures.

Although our results seem convincing, it is important to recognise that they have limitations:

Data Availability and Quality: The completeness and quality of the health and AQI data we used will determine how accurate our results will be. Inconsistent data or scope constraints may distort the results and the models' ability to forecast the future.

Regional Variability: Due to the considerable regional variations in health infrastructure and the sources of air pollution, the study's relevance may be restricted.

Future Directions for Research

Building on the findings of this study, more research may go in a number of directions:

Broader Health Outcomes: A more thorough knowledge of the effects of air pollution might be obtained by expanding the research to include other health outcomes connected to it, such as cardiovascular disorders.

Long-run Impact Studies: Analysing how persistent improvements in air quality affect lung cancer and other health outcomes over the long run would provide important information on how successful pollution management strategies are.

This study emphasises how critical it is to develop integrated strategies that successfully address the problems caused by air pollution by fusing environmental science, public health advocacy, and policy-making.

CHAPTER 5

Summary, Conclusion, Recommendation, and Implication for Future Research

5.1 Summary of Findings

This research embarked on a detailed exploration of the link between the Air Quality Index (AQI) and lung cancer mortality across various regions in the United States. Utilizing a comprehensive dataset that includes a range of contaminants, this study employed advanced statistical and machine learning approaches to dig deep into the predicted associations. Key findings demonstrated that particulate matter (PM2.5 and PM10) and nitrogen dioxide (NO2) are significant predictors of lung cancer mortality, agreeing with current research that indicates to their harmful health implications. Through the use of models such as XGBoost and Artificial Neural Networks (ANN), the study revealed the capacity to anticipate lung cancer mortality with noteworthy accuracy, underlining the strength and promise of machine learning in environmental health investigations.

5.2 Conclusions

The findings reported in this study indisputably shows a robust and statistically significant link between high AQI and higher risks of lung cancer death. The impact of various contaminants in aggravating health risks underlines the urgent need for tailored public health responses and legislative actions. The efficacy of predictive modeling in this context highlights its crucial applicability in preventative health management and policy formation, offering a trustworthy tool for public health authorities and politicians.

5.3 Implications for Future Research

The findings suggests numerous options for future inquiry that might broaden and increase the understanding of air pollution’s larger implications:

Expansion to Broader Health Outcomes: Future research should study the influence of air quality on a wider array of health outcomes, including but not limited to cardiovascular illnesses and chronic respiratory problems, to give a comprehensive understanding of its health implications.

Longitudinal Impact Studies: There is a considerable gap in longitudinal evidence about the long-term health implications of continuous exposure to poor air quality. Longitudinal research might give insight on the chronic impacts and perhaps show delayed implications of exposure to contaminants.

Advanced prediction Models: While models utilised in this work gave substantial insights, there is space for the inclusion of more advanced machine learning algorithms and ensemble approaches that might further increase the accuracy and prediction capacity of these tools.

5.4 Recommendations

Based on the findings of this study, the following strategic suggestions are proposed:

Stricter Air Quality Regulations: Policies targeted at lowering emissions from important sources such as industrial operations and vehicle traffic should be reinforced. The implementation of higher air quality rules might minimise the amounts of dangerous pollutants dramatically.

Enhanced Public Health Initiatives: It is vital to create more public knowledge of the hazards linked with air pollution. Health programmes should focus on preventative measures and advocate steps that people and communities may do to decrease their exposure to dangerous contaminants.

5.5 Final Thoughts

This research underlines the essential relationship between environmental variables such as air quality and public health effects. The findings call for a multidisciplinary strategy that blends scientific research, public health policy, and community participation to overcome the difficulties caused by air pollution. Effective management and major increase in air quality can lead to significant gains in public health and welfare, notably for lung cancer mortality. The paper recommends concentrated efforts towards tougher environmental legislation, comprehensive public health frameworks, and persistent research to understand the complexity of air quality and health.

Reference

1. World Health Organization. (2021). Ambient (outdoor) air quality and health. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
2. Environmental Protection Agency. (2021). Air Quality Index (AQI) Basics. https://www.epa.gov/
3. Brunekreef, B., Strak, M., Chen, J., Andersen, Z. J., Atkinson, R., Bauwelinck, M., ... & Hoek, G. (2021). Mortality and morbidity effects of long-term exposure to low-level PM2. 5, bc, NO2, and O3: an analysis of European cohorts in the ELAPSE project. *Research Reports: Health Effects Institute*, *2021*.
4. World's Air Pollution: Real-time Air Quality Index. https://waqi.info/#/c/40.076/-94.277/5z
5. American Lung Association - State of the Air. https://www.lung.org/research/sota
6. Nieto, P. G., Combarro, E. F., del Coz Díaz, J. J., & Montañés, E. (2013). A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. *Applied Mathematics and Computation*, *219*(17), 8923-8937.
7. Cakir, S., & Sita, M. (2020). Evaluating the performance of ANN in predicting the concentrations of ambient air pollutants in Nicosia. *Atmospheric Pollution Research*, *11*(12), 2327-2334.
8. Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*, *14*(1), 6795.
9. Ji, J. S., Liu, L., Zhang, J., Kan, H., Zhao, B., Burkart, K. G., & Zeng, Y. (2022). NO2 and PM2. 5 air pollution co-exposure and temperature effect modification on pre-mature mortality in advanced age: a longitudinal cohort study in China. *Environmental Health*, *21*(1), 97.
10. Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*, *14*(1), 6795.
11. Sumathi, K., Aishwarya, K., Gayathri, M., Gayathri, R., & Menaha, A. (2021). Air quality prediction using classification techniques. *Annals of the Romanian Society for Cell Biology*, 3794-3805.
12. Maltare, N. N., & Vahora, S. (2023). Air Quality Index prediction using machine learning for Ahmedabad city. *Digital Chemical Engineering*, *7*, 100093.
13. Hossain, I., Rahman, M. S., Sattar, S., Haque, M., Mullick, A. R., Siraj, S., ... & Khan, M. H. (2021). Environmental overview of air quality index (AQI) in Bangladesh: Characteristics and challenges in present era. *International Journal of Research in Engineering, Science and Management*, *4*(7), 110-115.
14. Soh, P. W., Chang, J. W., & Huang, J. W. (2018). Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. *Ieee Access*, *6*, 38186-38199.

15. Alam, M. Z., Armin, E., Haque, M., Kayesh, J. H. E., & Qayum, A. (2018). Air pollutants and their possible health effects at different locations in Dhaka City. *Journal of Current Chemical Pharmaceutical Sciences*, *8*(1), 111.
16. TY - BOOK AU - Rana, Masud AU - Biswas, Swapan PY - 2019/01/21 SP - T1 - Ambient air quality in Bangladesh 2012 - 2018 VL - DO - 10.13140/RG.2.2.14741.17128 ER -
17. Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B. P., Mohanavel, V., ... & Asfaw, A. K. (2022). [Retracted] Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International*, *2022*(1), 6254177.
18. Sumathi, K., Aishwarya, K., Gayathri, M., Gayathri, R., & Menaha, A. (2021). Air quality prediction using classification techniques. *Annals of the Romanian Society for Cell Biology*, 3794-3805.
19. Kamis, A., Cao, R., He, Y., Tian, Y., & Wu, C. (2021). Predicting lung cancer in the United States: A multiple model examination of public health factors. *International journal of environmental research and public health*, *18*(11), 6127.
20. Bhattacharya, S., & Shahnawaz, S. (2021). Using machine learning to predict air quality index in new delhi. *arXiv preprint arXiv:2112.05753*.
21. Torre P Rd, Zeldow B, Yao TJ, Hoffman HJ, Siberry GK, Purswani MU, Frederick T, Spector SA, Williams PL. Newborn Hearing Screenings in Human Immunodeficiency Virus-Exposed Uninfected Infants. J AIDS Immune Res. 2016;1(1):102. Epub 2016 Sep 5. PMID: 28459118; PMCID: PMC5407375.
22. Leading Causes of Death. 2021. Available online: https://cdc.gov/nchs/fastats/leading-causes-of-death.htm (accessed on 3 June 2024).
23. Pope Iii, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, *287*(9), 1132-1141.
24. Jacobson, M. Z. (2008). On the causal link between carbon dioxide and air pollution mortality. *Geophysical Research Letters*, *35*(3).
25. Anenberg, S. C., Horowitz, L. W., Tong, D. Q., & West, J. J. (2010). An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. *Environmental health perspectives*, *118*(9), 1189-1195.
26. Singh, G. K., Williams, S. D., Siahpush, M., & Mulhollen, A. (2011). Socioeconomic, rural-urban, and racial inequalities in US cancer mortality: part I—all cancers and lung cancer and part II—colorectal, prostate, breast, and cervical cancers. *Journal of cancer epidemiology*, *2011*(1), 107497.
27. Gharibvand, L., Shavlik, D., Ghamsary, M., Beeson, W. L., Soret, S., Knutsen, R., & Knutsen, S. F. (2017). The association between ambient fine particulate air pollution and lung cancer incidence: results from the AHSMOG-2 study. *Environmental health perspectives*, *125*(3), 378-384.

Chapters (2).docx

ORIGINALITY REPORT

13%	11%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	worldwidescience.org Internet Source	1%
3	Nilesh N. Maltare, Safvan Vahora. "Air Quality Index prediction using machine learning for Ahmedabad city", Digital Chemical Engineering, 2023 Publication	1%
4	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
5	Abdelaziz Testas. "Distributed Machine Learning with PySpark", Springer Science and Business Media LLC, 2023 Publication	<1%
6	www.researchgate.net Internet Source	<1%
7	ia902502.us.archive.org Internet Source	<1%