University of Nairobi

Department of Mathematics
Faculty of Science and Technology

PREDICTING RESPIRATORY DISEASES ATTRIBUTED TO PM2.5 AIR POLLUTION IN NAIROBI COUNTY, KENYA USING RANDOM FOREST

BY

VALINE OKEYO

THIS PROJECT IS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF MASTER OF PUBLIC HEALTH DATA SCIENCE OF THE UNIVERSITY OF NAIROBI

July 10, 2024

Contents

Declaration

Declaration

Researcher's Declaration

This project report is my original work and has not been presented in any other institution for any academic award. SIGNATURE: DATE: 09/07/2024

Valine Okeyo

Registration Number: SDS6/43905/2023

Supervisor's Approval

This project has been submitted in partial fulfillment of the requirements for the Degree of Master in Public Health Data Science of the University of Nairobi with my approval as the university supervisor. SIGNATURE: DATE: 09/07/2024

Dr. Idah Orowe

Faculty of Science and Technology
University of Nairobi

Dedication

Dedication

This work is dedicated to my husband Austine Okoth, and our son Jonathan Okoth.

Acknowledgement

Acknowledgement

I express my deepest gratitude to the Almighty for His abundant grace throughout this project. I extend my heartfelt thanks to my supervisor, Dr. Idah Orowe, for her unwavering support and guidance, which were instrumental to the success of this research. I am also grateful to Prof. Nicholas Oguge and the entire staff of the GEO-Health Hub at the University of Nairobi for their invaluable assistance with the data necessary for this study. Additionally, I would like to acknowledge the support of my family and friends. To my husband, Austine Okoth, father, David Okeyo, my mother, Pamela Okeyo, and my uncle, Elisha Odira, your encouragement, and assistance have been greatly appreciated. Thank you all for your contributions and unwavering support.

List of Tables

List of Figures

List of Abbreviations

ML	Machine Learning
EHR	Electronic Health Records
WHO	World Health Organization
PM2.5	Particulate Matter 2.5
COPD	Chronic Obstructive Pulmonary Disease
ARIs	Acute Respiratory Infections
URI	Upper Respiratory Infections
RF	Random Forest
ROSE	Random Oversampling Examples
BAM	Beta Attenuated Monitor
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
NN	Neural Network
MLP	Multilayer Perceptron
SVM	Support Vector Machine
GBM	Gradient Boosting Machine
GPR	Gaussian Process Regression
MAE	Mean Absolute Error
MSE	Mean Squared Error
SVR	Support Vector Regression
SHAP	Shapley Additive Explanation

Definition of Terms

Machine Learning Machine learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit instructions. Instead, these models learn patterns from data and make predictions or decisions based on that learning.

Random Forest Random Forest is a machine learning algorithm known for its versatility and robustness. It belongs to the ensemble learning family, which means it combines multiple algorithms to improve performance.

Respiratory Diseases Respiratory diseases are medical conditions that affect the airways and other structures of the lungs. These diseases can range from mild conditions, such as the common cold, to severe illnesses like chronic obstructive pulmonary disease (COPD), asthma, pneumonia, and lung cancer. They can impair lung function, causing difficulty in breathing and reducing the efficiency of oxygen exchange in the body.

PM2.5 PM2.5 refers to fine particulate matter that is less than 2.5 micrometers in diameter. These tiny particles can come from various sources, including vehicle emissions, industrial processes, combustion, and natural sources like wildfires.

Abstract

This study investigates the predictive capability of a Random Forest model in identifying respiratory diseases attributed to PM2.5 exposure in Nairobi County. Leveraging a comprehensive dataset encompassing demographic and air quality variables, the model demonstrated robust performance metrics, achieving an accuracy of 79.97% and an area under the curve (AUC) of 0.872. These results highlight the model's effectiveness in distinguishing between respiratory and cardiovascular conditions. The model's sensitivity and specificity were 81.88% and 73.27%, respectively, indicating a strong ability to correctly identify both true positives and true negatives. Analysis of feature importance revealed that age and PM2.5 concentrations were the most influential factors in predicting health outcomes, emphasizing the significant impact of air pollution and demographic factors on respiratory and cardiovascular health. Furthermore, the consistent train and test error rates across varying training set sizes suggest the model's stability and generalizability. This study underscores the importance of addressing air quality issues to mitigate the health impacts of PM2.5 exposure in urban settings.

Keywords: Respiratory diseases, PM2.5, Machine Learning, Random Forest, AUC, Sensitivity, Specificity, Feature Importance.

Chapter 1

INTRODUCTION

1.1 Background to the Study

Air pollution, particularly fine particulate matter (PM2.5), is a critical environmental and public health concern worldwide. PM2.5, consisting of tiny particles with a diameter of 2.5 micrometers or smaller, poses significant health risks due to its ability to penetrate deep into the respiratory tract. Exposure to PM2.5 is associated with a range of adverse health outcomes, including respiratory diseases such as asthma, bronchitis, chronic obstructive pulmonary disease (COPD), and lung cancer. The World Health Organization (WHO) attributes millions of premature deaths annually to air pollution, with a substantial proportion linked to PM2.5 exposure (WHO, 2018).

Numerous studies have established the link between PM2.5 and respiratory diseases. For instance, Pope et al. (2002) demonstrated that long-term exposure to PM2.5 increases the risk of cardiopulmonary mortality and lung cancer. Similarly, a study by Anderson et al. (2012) highlighted the association between short-term PM2.5 exposure and exacerbation of asthma and COPD symptoms. These studies underscore the global significance of PM2.5 as a major health hazard, necessitating localized research to understand its impacts in specific urban contexts.

Nairobi, the capital city of Kenya, is undergoing rapid urbanization and industrialization, contributing to worsening air quality. The city's population has surged in recent decades, leading to increased motor vehicle emissions, construction activities, and industrial operations. Additionally, many residents in informal settlements rely on biomass and kerosene for cooking and heating, further exacerbating air pollution levels (Gatari et al., 2015). Consequently, Nairobi's residents are frequently exposed to harmful concentrations of PM2.5, raising serious public health concerns.

The health impacts of air pollution in Nairobi are particularly concerning given the city's limited healthcare infrastructure and resources. Respiratory diseases are already a significant burden, and the additional strain from pollution-related health issues poses a challenge to the healthcare system. Vulnerable populations, including children, the elderly, and individuals with pre-existing health conditions, are at a higher risk of adverse effects from PM2.5 exposure (Egondi et al., 2018). Despite these challenges, there is a noticeable gap in local research focusing on the specific relationship between PM2.5 levels and respiratory health outcomes in Nairobi.

Current air quality monitoring efforts in Nairobi are fragmented and often lack the granularity needed for comprehensive analysis. Although some air quality monitoring stations exist, data collection is inconsistent and not well-integrated with health surveillance systems (Onyango et al., 2021). This limitation hinders detailed epidemiological studies and the development of effective interventions. Additionally, public awareness about the health risks associated with air pollution remains low, limiting the community's ability to take precautionary measures.

Machine learning techniques, particularly the Random Forest algorithm, offer a promising approach to addressing these challenges. The Random Forest algorithm is a versatile and powerful tool for predictive modeling, capable of handling complex datasets with numerous variables. It works by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks or the mean prediction for regression tasks. This ensemble method helps to improve predictive accuracy and reduce the risk of overfitting (Breiman, 2001).

In the context of predicting respiratory diseases attributed to PM2.5 in Nairobi, the Random Forest algorithm can analyze various data sources, including air quality indices, meteorological data, health records, and demographic information. By identifying patterns and relationships within these datasets, the algorithm can predict the likelihood of respiratory disease occurrences based on PM2.5 exposure levels. This predictive capability can provide valuable insights for public health officials, enabling them to implement targeted interventions and allocate resources more effectively (Ravindra et al., 2023).

The integration of predictive modeling into public health strategies represents a significant advancement in addressing air pollution-related diseases. Studies such as those by Li et al. (2019) and Patel et al. (2018) have demonstrated the efficacy of machine learning models, including Random Forest, in predicting health outcomes and identifying high-risk patients. By developing a robust predictive model for Nairobi, this study aims to fill the existing research gap and contribute to the body of knowledge on air pollution and health.

1.2 Statement of the Problem

Despite global awareness of the health risks posed by PM2.5 air pollution, localized predictive studies focusing on Nairobi are scarce. This lack of localized research creates a significant gap in understanding and addressing the public health challenges associated with air pollution in the city.

Nairobi's rapid urbanization, increased vehicular emissions, and industrial activities have significantly deteriorated air quality, contributing to higher incidences of respiratory diseases. Current air quality monitoring and management strategies in Nairobi are fragmented and insufficient, resulting in sporadic and poorly integrated data on PM2.5 levels and health outcomes. This fragmented approach hinders the ability to draw concrete correlations between pollution levels and respiratory health, posing a challenge for public health interventions.

Moreover, public awareness regarding the impact of air pollution on respiratory health remains low. This is exacerbated by the lack of accessible, actionable information, which limits the community's ability to take preventive measures.

To address these challenges, there is an urgent need to develop a robust predictive model. Such a model can bridge existing gaps by providing timely insights into potential health risks and enabling proactive measures. The Random Forest algorithm, known for its accuracy and ability to handle complex datasets, offers a promising solution for predicting respiratory diseases related to PM2.5 exposure. By analyzing large volumes of heterogeneous data, this algorithm can forecast disease occurrences and provide a comprehensive understanding of the contributing factors.

This study aims to fill the existing research gap by developing a predictive model tailored to Nairobi's unique context. The model's outputs can inform public health strategies, such as issuing health advisories, optimizing healthcare resource allocation, and designing targeted interventions for vulnerable populations. Ultimately, this project seeks to contribute to the scientific community and provide practical benefits for Nairobi's residents, promoting a healthier and more sustainable urban environment.

1.3 General Objective

The general objective of this study is to develop a predictive model using the Random Forest algorithm to forecast respiratory diseases attributed to PM2.5 exposure in Nairobi, Kenya.

1.4 Specific Objectives

1. Develop a predictive machine learning model using Random Forest to forecast the occurrence of respiratory diseases in Nairobi County using historical health data and PM2.5 air pollution levels.

2. Examine the feature importance rankings provided by the Random Forest algorithm to gain insights into the relative contributions of different variables in predicting respiratory disease outcomes.

3. Evaluate the performance and accuracy of the Random Forest predictive model in predicting respiratory disease occurrences attributed to PM2.5 air pollutants in Nairobi County.

1.5 Research Questions

1. How can a predictive machine learning model be developed using the Random Forest algorithm to forecast the occurrence of respiratory diseases attributed to PM2.5 in Nairobi County, Kenya?

2. What insights can be gained from the feature importance rankings provided by the Random Forest algorithm regarding the relative contributions of different variables in predicting respiratory disease outcomes in Nairobi County?

3. What is the performance and accuracy of the Random Forest predictive model in predicting respiratory disease occurrences attributed to PM2.5 air pollutants in Nairobi County, Kenya?

1.6 Significance of the Study

The utilization of machine learning techniques for predictive analytics in the context of respiratory diseases and PM2.5 air pollution represents a novel approach to public health research. The study can contribute to the advancement of methodologies in this field, offering insights into the application of cutting-edge technologies for addressing complex health-related challenges.

By demonstrating the effectiveness of machine learning algorithms in predicting respiratory disease outbreaks in relation to PM2.5 air pollution levels, the study can contribute to evidence-based environmental policy-making. This may lead to the development and implementation of stricter air quality regulations and initiatives aimed at reducing pollution levels, thereby promoting a healthier environment for residents of Nairobi County.

Understanding the relationship between PM2.5 air pollution and respiratory diseases through predictive analytics can provide valuable insights into the factors contributing to public health issues in Nairobi County. This knowledge can inform policymakers and healthcare professionals in implementing targeted interventions to reduce the burden of respiratory illnesses and improve overall health outcomes among the population.

By raising awareness of the impacts of PM2.5 air pollution on respiratory health and providing predictive analytics tools, the study can empower individuals and communities to take proactive measures to protect themselves from the adverse effects of poor air quality. This may include advocating for cleaner energy sources, promoting sustainable transportation options, and adopting personal protective measures during periods of high pollution levels.

1.7 Scope of the Study

This study focuses on Nairobi County, Kenya, examining the occurrence and severity of respiratory diseases within this geographic area. The study utilizes historical health data, including records of respiratory disease cases, as well as data on PM2.5 air pollution levels obtained from relevant monitoring stations within Nairobi County. The Random Forest machine learning algorithm is employed to analyze the relationship between air pollution levels and respiratory disease occurrences.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This literature review synthesizes current knowledge on the impacts of PM2.5 air pollution on respiratory health, encompassing epidemiological studies, health outcomes, risk factors, and mitigation strategies. The review aims to provide a comprehensive understanding of the complex interactions between PM2.5 exposure and respiratory diseases across different populations and geographic regions. In recent years, a variety of studies have employed different machine learning (ML) models to predict respiratory diseases and assess their relationships with air pollution. This section also reviews key studies, their objectives, methodologies, comparative results with Random Forest (RF), and discusses their limitations.

2.2 Epidemiological Evidence Linking PM2.5 to Respiratory Diseases

Epidemiological research over the past decades has consistently demonstrated the adverse effects of PM2.5 exposure on respiratory health. Fine particulate matter, defined as particles with a diameter of 2.5 micrometers or less, penetrates deep into the respiratory tract upon inhalation, reaching the alveoli and potentially triggering inflammation and oxidative stress (Dockery et al., 1993). Studies have shown that acute and chronic exposure to PM2.5 is associated with increased incidence and exacerbation of respiratory conditions such as asthma, bronchitis, chronic obstructive pulmonary disease (COPD), and respiratory infections (Pope et al., 2002; HEI, 2010).

For instance, Dockery et al. (1993) conducted pioneering research linking short-term exposure to PM2.5 and other particulate matter with acute respiratory symptoms and exacerbations of pre-existing respiratory diseases. This early evidence laid the foundation for subsequent studies exploring the long-term health impacts of PM2.5 exposure. Pope et al. (2002) further demonstrated the association between long-term exposure to fine particulate air pollution and increased mortality from lung cancer, cardiovascular diseases, and respiratory ailments, highlighting the chronic health risks posed by PM2.5.

In Sub-Saharan Africa (SSA), rapid urbanization, industrialization, and population growth have contributed to escalating levels of air pollution, exacerbating respiratory health risks in urban centers (Abera et al., 2021). Air pollutants such as PM2.5, carbon monoxide (CO), sulfur dioxide (SO2), nitrogen dioxide (NO2), and ozone (O3) often exceed WHO guidelines, posing significant public health challenges (Agbo et al., 2021).

Research specific to SSA highlights the impact of biomass burning for cooking and heating on indoor air quality and respiratory health. Kurmi et al. (2012) found that biomass cooking fuels are associated with increased incidence of acute respiratory infections (ARI) among children, contributing to high childhood mortality rates in rural and peri-urban areas. Moreover, Egondi et al. (2018) documented elevated risks of respiratory symptoms and diseases among children exposed to high levels of ambient PM2.5 in urban areas, emphasizing the need for targeted interventions to improve air quality and protect vulnerable populations.

In Asia, particularly in rapidly developing countries like China and India, PM2.5 pollution has reached alarming levels due to industrial emissions, vehicular exhaust, and biomass burning (Zeng et al., 2016). Studies have linked prolonged exposure to high levels of PM2.5 with increased respiratory morbidity and mortality, highlighting the urgent need for stringent air quality regulations and sustainable development practices (HEI, 2010; Cui et al., 2019).

Research in South Korea by Habre et al. (2014) demonstrated a clear association between outdoor PM2.5 exposure and respiratory symptoms in children with asthma, underscoring the immediate health impacts of air pollution on vulnerable populations. Similarly, studies in Japan have shown dose-response relationships between PM2.5 exposure and respiratory health outcomes, reinforcing the importance of air quality management in densely populated urban areas (Higashi et al., 2014).

In Europe and North America, extensive air quality monitoring and research have established robust links between PM2.5 exposure and adverse respiratory health outcomes (EEA, 2020). Cohort studies have consistently reported increased hospital admissions for respiratory conditions, exacerbations of asthma and COPD, and higher mortality rates associated with long-term exposure to PM2.5 (Burnett et al., 2018; Anderson et al., 2012).

For example, research in the United States has highlighted disparities in PM2.5 exposure and health outcomes among socioeconomically disadvantaged communities, where higher pollution levels exacerbate respiratory health disparities (Spira-Cohen et al., 2011). These findings underscore the need for targeted public health interventions and policies to reduce air pollution and mitigate its health impacts on vulnerable populations.

Studies in Europe have associated long-term exposure to nitrogen monoxide (NO) and PM with increased incidence of rhinitis and other respiratory symptoms (Burte et al., 2020). Similarly, research from South Korea has linked higher exposure to outdoor PM2.5 components with increased coughing and wheezing in children with asthma (Habre et al., 2014).

Children living near electronic waste recycling sites in China, where PM2.5 levels are elevated, face heightened risks of respiratory symptoms such as coughing compared to children in less polluted areas (Zeng et al., 2016). Studies in Japan have identified a dose-response relationship between delayed PM2.5 exposure and coughing in both asthmatic and non-asthmatic individuals, highlighting the widespread impact of particulate pollution on respiratory health (Higashi et al., 2014).

The pathophysiological mechanisms underlying the respiratory effects of PM2.5 involve oxidative stress, inflammation, and impaired lung function. Fine particulate matter can penetrate deep into the lungs, triggering inflammatory responses in lung tissues and systemic oxidative stress that contributes to cardiovascular and respiratory diseases (Kelly & Fussell, 2015). The deposition of PM2.5 in the alveoli and airways can lead to epithelial cell damage, activation of inflammatory pathways, and increased susceptibility to respiratory infections (Li et al., 2017).

Moreover, PM2.5 exposure has been linked to systemic effects beyond the respiratory system, including cardiovascular diseases, neurodevelopmental disorders, and adverse pregnancy outcomes, highlighting the multifaceted health risks associated with air pollution (Brook et al., 2010; Block & Calderón-Garciduenas, 2009).

Effective mitigation of PM2.5 pollution requires integrated approaches combining regulatory measures, technological innovations, and public awareness campaigns. Policy interventions such as emission standards for vehicles and industrial sources, promotion of cleaner technologies, and urban planning initiatives to reduce traffic congestion and promote green spaces are crucial for improving air quality (Gupta et al., 2012).

Internationally, initiatives such as the WHO Air Quality Guidelines and regional agreements on air pollution control play pivotal roles in setting standards and guiding countries towards sustainable development practices that prioritize public health (WHO, 2021). Collaborative efforts between governments, researchers, and civil society are essential for implementing and monitoring these policies to ensure effective reduction of PM2.5 levels and protection of respiratory health worldwide.

PM2.5 air pollution poses significant risks to respiratory health globally, impacting diverse populations across different geographic regions. Epidemiological evidence consistently links PM2.5 exposure to increased morbidity and mortality from respiratory diseases, underscoring the urgent need for comprehensive air quality management strategies. Addressing the complex challenges of air pollution requires interdisciplinary research, evidence-based policies, and international cooperation to safeguard public health and promote sustainable development.

2.3 Predictive Modeling with Machine Learning

In recent years, machine learning algorithms, particularly Random Forest (RF), have gained prominence in environmental health research for their ability to model complex relationships and predict PM2.5 concentrations with high accuracy. RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks or the mean prediction for regression tasks. This approach is particularly advantageous in environmental studies where variables exhibit non-linear relationships and interactions.

Martínez et al. (2018) and Kaminska (2018) have highlighted RF as a preferred method for predicting fine particulate matter due to its capability to handle incomplete datasets and discern non-linear relationships among data points. The ensemble nature of RF enables it to capture complex interactions between meteorological variables, emissions sources, and geographical factors influencing PM2.5 concentrations. By aggregating predictions from multiple decision trees, RF enhances predictive accuracy and robustness in air quality modeling (Martínez et al., 2018).

Jing et al. (2023) employed Support Vector Regression (SVR) to forecast hospital visits for respiratory diseases (pneumonia, acute upper respiratory infections, chronic lower respiratory diseases) based on meteorological and air pollution data in Linyi, China. Their study demonstrated moderate to high prediction accuracy, with SVR effectively modeling the impact of environmental factors on respiratory health outcomes. However, the model demonstrated the highest prediction accuracy for pneumonia among the three types of respiratory diseases. Therefore, there is a need to augment the sample size of health data and develop a robust machine learning model, such as Random Forest, capable of effectively generalizing across all respiratory diseases.

Ravindra et al. (2023) utilized machine learning techniques to forecast the impact of ambient air pollution on outpatient visits for acute respiratory infections (ARI) in India. They employed eight ML algorithms, including Random Forest, and demonstrated RF's superior performance with high R^2 values for predicting ARI-related hospital visits.

Yunseo et al. (2022) utilized machine learning models to forecast respiratory diseases based on climatic and air pollution variables in Seoul, South Korea. They collected daily data on respiratory disease patients and employed a relief-based feature selection algorithm to determine significant features. Two predictive models were developed: gradient boosting and Gaussian process regression (GPR). Gradient boosting iteratively improved prediction accuracy by fitting residuals, achieving an R^2 of 0.68 and RMSE of 13.8 on unseen test data. GPR, known for time series prediction, achieved an R^2 of 0.67 and RMSE of 13.9. SHAP analysis 14 identified atmospheric pressure, carbon monoxide (CO), and PM2.5 as influential factors. The main limitation of this study, however, was that daily patient data excluded factors like demographics and medical history which would have significantly contributed to the performance of the model and insight examination.

Li et al. (2021) compared Gradient Boosting Machines (GBM) and Random Forest (RF) for predicting respiratory morbidity rates associated with PM2.5 exposure across different regions. Both models showed comparable accuracy, with GBM highlighting regional variability in respiratory health outcomes compared to RF.

Zhao et al. (2020) evaluated Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN) for predicting hospital admissions due to respiratory diseases in polluted urban environments. RF demonstrated superior performance in capturing non-linear relationships and handling high-dimensional data compared to SVM and NN.

Li Luo et al. (2019) employed machine learning techniques to predict high-cost patients and investigate key variables using medical insurance data from a large city in western China. Logistic regression and Random Forest (RF) models were compared, with both demonstrating good predictive performance. The RF model achieved an area under the ROC curve (AUC) of 0.801, outperforming logistic regression slightly, which achieved an AUC of 0.787. The study highlighted the effectiveness of Random Forest in healthcare cost prediction and identified variables crucial for accurate modeling.

Patel et al. (2018) conducted a retrospective analysis of pediatric asthma exacerbations in London, UK, using decision trees, LASSO logistic regression, Random Forest, and Gradient Boosting Machines (GBM). The study found that all models performed well in predicting hospitalization risk, with AUC values ranging 15 from 0.72 (decision trees) to 0.84 (GBM). Gradient boosting showed a slight advantage in predicting the need for hospital-level care, emphasizing its utility in triaging pediatric patients with asthma exacerbations. The main limitation of this study is that it was performed at a single institution.

Mei-Juan Chen et al. (2018) utilized machine learning to correlate PM2.5 and PM10 concentrations with outpatient visits for upper respiratory tract infections (URIs) in Taiwan. They employed a Multilayer Perceptron (MLP) to predict URI volumes based on PM concentrations. The study revealed seasonal variations in PM concentrations and URI cases, with MLP achieving high accuracy (up to 89.05

Dimitris et al. (2017) investigated clinical decision support systems for respiratory diseases like asthma and COPD using machine learning classifiers. They compared seven classifiers, including Random Forest, across 132 patients in Greece. Random Forest outperformed other methods, achieving high precision (97.716 diagnosing respiratory diseases and emphasized the importance of specific clinical variables in accurate prediction models.

Study	**Model**	**Performance (R^2)**
Patel et al. (2018)	Decision Trees	0.72
Patel et al. (2018)	Random Forest	0.84
Jing et al. (2023)	Support Vector Regression (SVR)	High
Ravindra et al. (2023)	Random Forest	High
Yunseo et al. (2022)	Gradient Boosting	0.68
Yunseo et al. (2022)	Gaussian Process Regression (GPR)	0.67
Li et al. (2021)	Gradient Boosting Machines (GBM)	Comparable
Zhao et al. (2020)	Support Vector Machines (SVM)	Inferior
Li Luo et al. (2019)	Logistic Regression	0.787
Mei-Juan Chen et al. (2018)	Multilayer Perceptron (MLP)	89.05%
Dimitris et al. (2017)	Random Forest	97.7% (COPD), 80.3% (Asthma

Table 2.1: Comparison of model performance in predicting respiratory diseases.

2.4 Comparative Analysis

These studies collectively demonstrate the diverse applications of machine learning in predicting and understanding respiratory diseases and their relationships with environmental factors. Random Forest emerges as a robust model in healthcare cost prediction (Li Luo et al., 2019) and diagnosis of respiratory conditions (Dimitris et al., 2017), Patel et al. (2018) and Mei-Juan Chen et al. (2018) explore its effectiveness alongside other models like Gradient Boosting and Multilayer Perceptron in predicting hospitalization risk and correlating PM concentrations with health outcomes.

2.5 Research Gap

While studies have developed prediction models for respiratory diseases in various countries, the applicability of these models to the unique context of Nairobi County, Kenya, remains uncertain. Additionally, existing statistical models may struggle to effectively analyze the vast amounts of data available in this setting. This research aims to bridge these gaps by utilizing Random Forest techniques to forecast respiratory diseases attributed to PM2.5 air pollution, leveraging historical health data and PM2.5 air pollution levels specifically within Nairobi County. By addressing these challenges, this study seeks to provide tailored insights into the effects of PM2.5 air pollutant on respiratory health in the Kenyan context, contributing to more accurate predictive analytics and informing targeted public health interventions.

2.6 Conceptual Framework

The conceptual framework for this study is grounded in the intersection of machine learning, public health, and environmental science. At its core, the framework acknowledges the intricate relationship between air pollution and respiratory diseases, particularly focusing on the PM2.5 air pollutant. Drawing upon established theories and models in epidemiology, environmental health, and machine learning, this study seeks to develop a predictive analytics model tailored to the unique context of Nairobi County, Kenya. Key components of the conceptual framework include the utilization of historical health data and PM2.5 air pollution levels for machine learning algorithms. The algorithm, selected based on its suitability for handling large datasets and predicting complex health outcomes, was trained and evaluated to forecast respiratory disease occurrences and severity. By integrating these multidisciplinary perspectives, the conceptual framework guides the systematic exploration of the impacts of PM2.5 air pollution on respiratory diseases in Nairobi County, facilitating a comprehensive understanding of the underlying mechanisms and informing targeted interventions to mitigate adverse health effects.

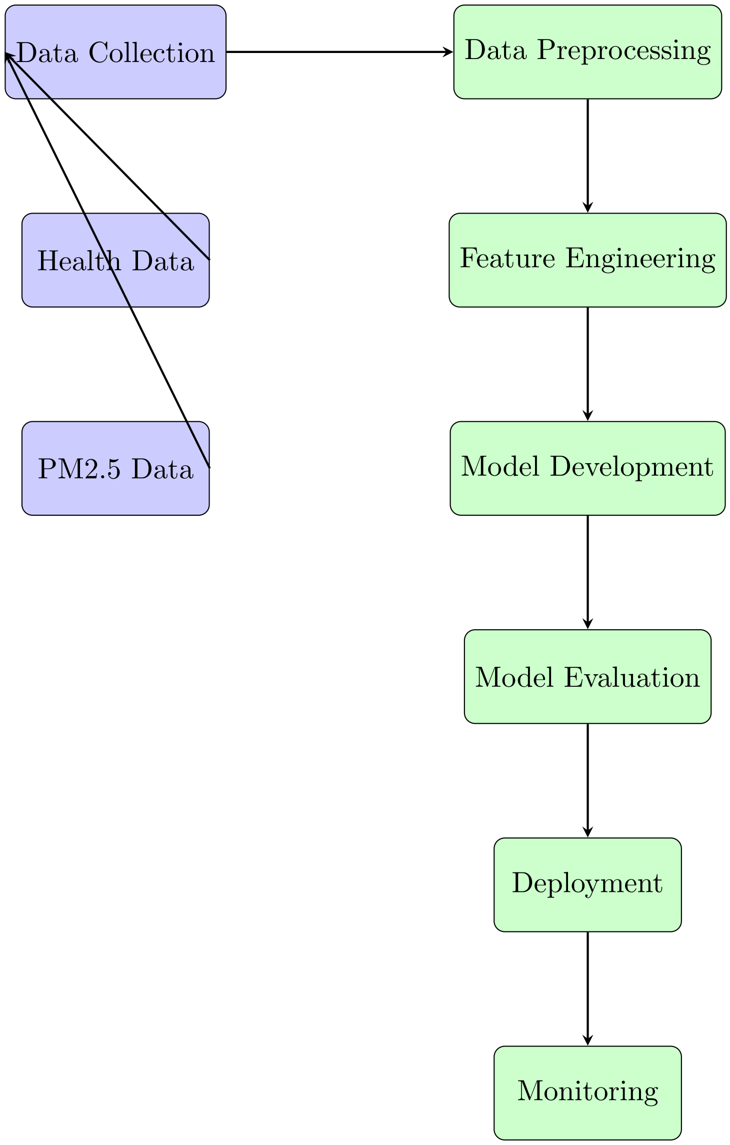

Figure 2.1: Conceptual Framework for Predicting Respiratory Diseases Attributed to PM2.5 Air Pollution

Chapter 3

RESEARCH METHODOLOGY

3.1 Introduction

This section outlines the methodological approach adopted to investigate the impact of PM2.5 air pollution on respiratory diseases in Nairobi County, Kenya. Central to this study is the utilization of the Random Forest algorithm to develop predictive models capable of forecasting respiratory disease incidences attributed to PM2.5 exposure. The methodology encompasses data collection, preprocessing, modeling techniques, and validation strategies aimed at generating robust insights into the health impacts of urban air pollution.

3.2 Research Site

Nairobi, the economic hub of Kenya, stands as the most populous city in East Africa and presently ranks as the 6th largest city on the African continent. According to the 2019 Census of Kenya, the administrative area of Nairobi city spans 6,317.6 km^2 (16,316 sq. mi) and was home to 4,750,056 inhabitants. The city is under the political administration of the county government and is divided into constituencies.

Health facilities in this city are categorized into two main classes: public and private health facilities. Public health facilities include health centers and hospitals owned by the Ministry of Health, while private facilities encompass healthcare institutions owned by private entities. Several factors influence the choice of this county as the study area, including the accessibility of health demand data and the range of health services offered at healthcare facilities. Additionally, the planning issues addressed in this county hold relevance for other hospitals throughout the country.

3.3 Research Design

To address the objectives of this study, both descriptive and analytical research designs were employed. Descriptive research was useful in data exploration. Analytical research was suitable for model development and evaluation.

3.4 Study Population

This study involved 2 national hospitals; private and public hospitals. The benchmark for selecting the hospitals in this study was based on the value of the data in question.

3.5 Sources of Data

The study obtained its data from the East Africa Global Environmental and Occupational Health Research and Training Center. This center served as the ideal source due to its comprehensive data collection from the aforementioned hospitals. Moreover, the center is equipped with a Beta Attenuation Monitor (BAM), facilitating the continuous, real-time measurement of PM2.5 concentrations.

3.6 Data Analysis

Data analysis played a pivotal role in this study, employing a range of analytical techniques to derive meaningful insights into the relationship between PM2.5 air pollution and respiratory diseases in Nairobi County, Kenya. The analytical framework encompasses several key approaches tailored to the study's objectives and dataset characteristics:

Descriptive Analytics Descriptive analytics forms the foundation of this study, involving the examination of historical data on PM2.5 concentrations and respiratory disease incidences. This phase aimed to summarize and interpret key trends, patterns, and distributions in the data over time and across different geographical areas within Nairobi County. By analyzing historical trends, descriptive analytics provides a comprehensive understanding of past changes and variations in air pollution levels and health outcomes.

Diagnostic Analytics Diagnostic analytics follows descriptive analytics and focuses on uncovering underlying causes and relationships between PM2.5 exposure levels and respiratory disease patterns. This phase utilized advanced statistical techniques and machine learning algorithms, such as MeanDecreaseGini analysis and Random Forest modeling, to identify significant associations and dependencies. By exploring causal relationships, diagnostic analytics seeks to elucidate how variations in PM2.5 concentrations contribute to respiratory health outcomes in Nairobi County.

Predictive Modeling with Random Forest Central to this study is the application of the Random Forest algorithm for predictive modeling. This machine learning technique is adept at handling complex datasets and capturing non-linear relationships between environmental exposures and health outcomes. The predictive model developed forecasted future incidences of respiratory diseases based on historical PM2.5 data, demographic factors, and other relevant variables. Evaluation metrics such as accuracy, sensitivity, and specificity were employed to assess the performance and reliability of the models.

Prescriptive Analytics Informed by the findings from descriptive, diagnostic, and predictive analyses, prescriptive insights were derived to inform targeted public health interventions. This phase of analysis aims to answer critical questions such as "What actions should be taken to mitigate the impact of PM2.5 air pollution on respiratory health?" Recommendations were tailored to the specific needs and challenges identified in Nairobi County, providing actionable strategies for policymakers and healthcare professionals.

Data Collection and Preparation

To achieve the objectives of this study, three-year data spanning from 2021 to 2023 of hospital data and PM2.5 data were collected. Monthly health records from various files were consolidated into a single folder. Similarly, daily PM2.5 records were gathered into another folder. These datasets were then individually imported into R, explored, and subsequently merged. Additionally, descriptive statistics was performed by examining the summary statistics of each variable and visualizing the data distributions to gain an initial understanding of the data.

Below are snippets of raw codes used in data preparation.

```
# Example of R code for data import and preparation
# Importing health data
health_data <- read.csv(”health_data.csv”)
# Importing PM2.5 data
pm_data <- read.csv(”pm25_data.csv”)
# Merging datasets
merged_data <- merge(health_data, pm_data, by=”date”)
# Descriptive statistics
summary(merged_data)
# Visualizing data
hist(merged_data$pm25, main=”Distribution␣of␣PM2.5”, xlab=”PM2.5␣Concentra
```

The health dataset had missing values in the ‘Appointment Date’, Age, and Sex columns. The missing appointment dates were imputed by the median date because the median is less affected by skewness and outliers than the mean, providing a more robust central tendency measure. Additionally, the median date represents the middle point of the data, ensuring that half the dates are before and half are after the imputed value. This helps maintain the chronological order and balance of the dataset.

Missing ages were imputed by the mean because the mean provides an average value that can fill in the gaps without distorting the overall age distribution significantly, especially because the number of missing values is relatively small compared to the dataset’s size.

Missing values in the ’Sex’ column were replaced by the letter ’U’ to represent ’Unknown’. The resulting summary statistics of the Age variable suggest a relatively young patient population, with half of the patients being 26 years or younger. The slightly higher mean of 27.79 suggests a slight right skew, indicating that while most patients are relatively young, some older patients increase the average age. The bar plot of the ’Sex’ variable indicates that females constitute the majority of the health records.

Summary Statistics of Age

- **Minimum:** 0.00 years
- **First Quartile (Q1):** 4 years
- **Median:** 26 years
- **Mean:** 27.79 years
- **Third Quartile (Q3):** 44 years
- **Maximum:** 100 years

Bar Plot of Gender

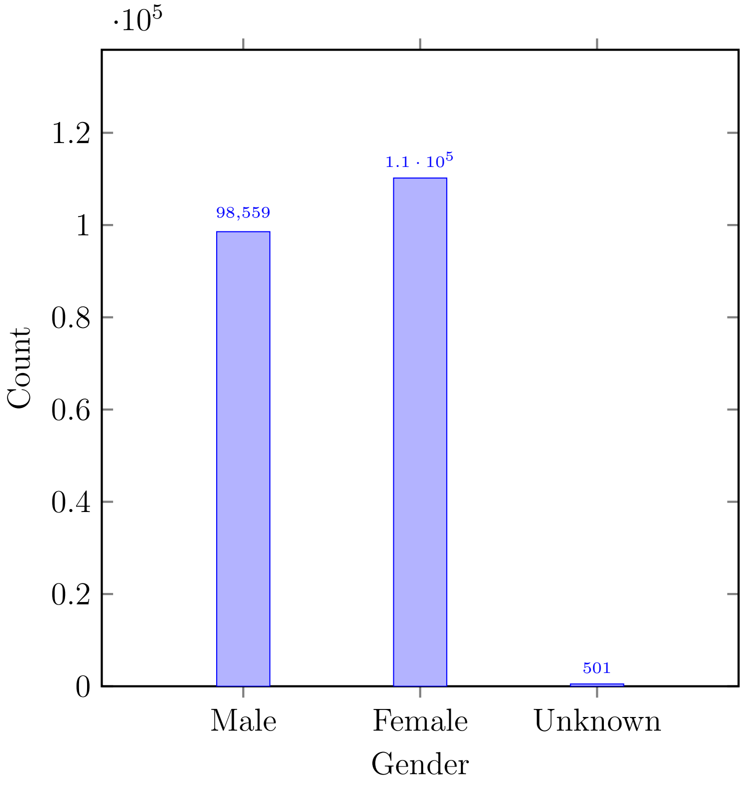

Figure 3.1: Gender Distribution

For PM2.5 data, the Concentration columns, Relative Humidity column, and status column were cleaned and adjusted to align with BAM (Beta Attenuation Monitor) calibrations. The summary statistics of the dataset reveal key insights into PM2.5 variables and environmental conditions. The concentration of PM2.5 ranges from a minimum of 3.00 µg/m^3 to a maximum of 108.00 µg/m^3, with a mean concentration of 20.88 µg/m^3 which exceeds the World Health Organization's recommended guideline for annual thresholds of 5 µg/m^3. Relative humidity spans from 11.00% to 99.00%, with a mean of approximately 70.02%. The Ambient temperature (AT.C.) varies between 16.40°C and 35.00°C, with a mean of 24.41°C. The summary also includes flow rate, BAM temperature (FT.C.), BAM humidity (FRH.), and other parameters such as barometric pressure (BP.mmHg). Additionally, the status column indicates the status of data points. Overall, the summary provides a comprehensive overview of PM2.5 variables and associated environmental factors within the dataset.

Histogram of PM2.5 Concentration

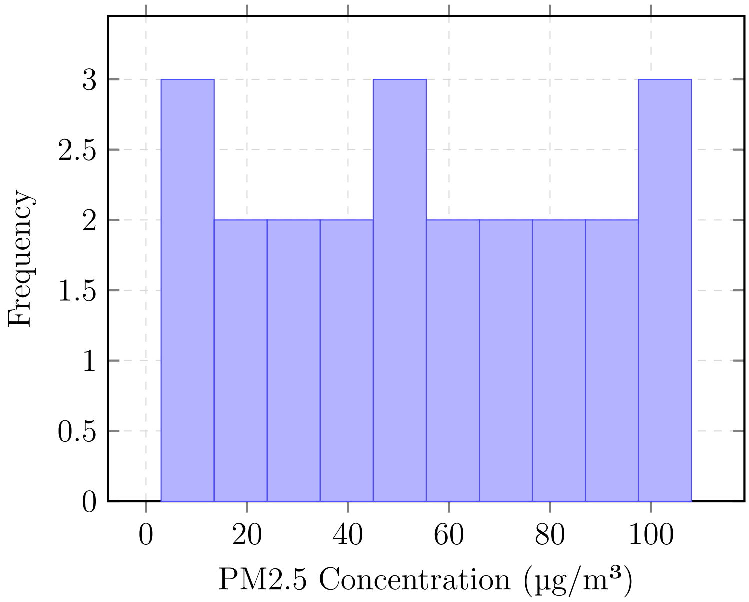

Figure 3.2: Histogram of PM2.5 Concentration Distribution

Feature Selection

Feature selection plays a significant role in improving the performance of machine learning algorithms by reducing the time to build the learning model and increasing the accuracy of the learning process. Out of the 16 features initially considered, only five features were selected to train the model: Real-time and Hourly $PM_{2.5}$ Concentrations, Age, Sex, and Diagnosis.

Model Selection

The choice of an appropriate model is pivotal in any data-driven study, influenced by factors such as dataset characteristics, the nature of the task at hand, and the specific types of data involved. In the realm of environmental health studies, where complexities abound in datasets featuring numerous variables and non-linear relationships, selecting an effective model is crucial for accurate prediction and insightful analysis.

For this study, the Random Forest algorithm was selected as the primary modeling approach for several compelling reasons:

- **High Accuracy and Robustness:** Random Forest is renowned for its ability to deliver high predictive accuracy and robust performance. This makes it particularly suitable for handling the intricate relationships between $PM_{2.5}$ air pollution levels and respiratory diseases within Nairobi County, Kenya.
- **Handling Large and Complex Datasets:** Environmental health datasets often encompass vast amounts of data points and numerous features. Random Forest excels in managing such large datasets efficiently, accommodating multiple predictors and interactions without compromising performance.

- **Efficient Handling of Missing Data:** Real-world datasets frequently contain missing values, which can hinder traditional modeling approaches. Random Forest is adept at managing missing data, ensuring reliable predictions even when data completeness is compromised.
- **Capturing Non-linear Relationships:** In environmental health contexts, relationships between pollutants and health outcomes are seldom linear. Random Forest's capability to capture non-linear relationships enables the exploration of complex interactions and their impacts on respiratory health.
- **Feature Importance Assessment:** One of the strengths of Random Forest lies in its ability to rank the importance of input features. This feature importance analysis aids in identifying the key factors driving respiratory diseases attributed to $PM_{2.5}$ air pollution, providing valuable insights for public health interventions.
- **Mitigation of Overfitting Risks:** Unlike individual decision trees, Random Forest mitigates the risk of overfitting due to its ensemble nature. This ensures that the model generalizes well to new data, maintaining robust performance across different scenarios.

Model Training

Given the complexity of environmental health datasets and the specific focus on Nairobi County, several considerations inform the model training process:

- **Data Imbalance:** Addressing potential class imbalances in respiratory disease data ensured that the model learns from a representative dataset, minimizing bias towards the majority class and improving overall predictive performance. Random Over-Sampling Examples (ROSE) technique was applied to randomly synthesize new examples by interpolating from the minority class to balance the class distribution.
- **Dataset Partitioning:** The dataset was partitioned into a 70% training set and a 30% test set to train the model on a sufficient amount of data and evaluate its performance on unseen data.
- **Validation Metrics:** Evaluation metrics such as accuracy, precision, recall, and F1-score were employed to assess the model's performance. Given the consequences of misclassifying respiratory diseases, these metrics provided insights into the model's ability to correctly identify and predict health outcomes.
- **Interpretability:** While Random Forest excels in predictive accuracy, efforts were made to interpret feature importance rankings derived from the model. This analysis elucidates which factors, such as $PM_{2.5}$ concentrations or demographic variables, exert the most significant influence on respiratory health outcomes in Nairobi County.

The feature importance of the variables used in training and testing the model were then examined using Mean Decrease Gini to gain insights into the factors influencing the target variable.

Feature	Mean Decrease Gini
ConcRT.ug.m^3	13217.7847
ConcHR.ug.m^3	13285.0345
AGE	45895.2029
SEX	838.4769

Table 3.1: Feature Importance Rankings using Mean Decrease Gini

Model Evaluation

Model evaluation is a crucial phase in assessing the performance and reliability of machine learning algorithms, ensuring that they generalize well to new, unseen data. In the context of this study on predicting respiratory diseases attributed to PM$_{2.5}$ air pollution using the Random Forest algorithm, rigorous evaluation metrics are employed to gauge the model's effectiveness. Key evaluation metrics include:

- **Accuracy:** Measures the overall correctness of predictions, indicating the proportion of correctly classified instances among the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

- **Precision and Recall:**
  - **Precision:** Measures the proportion of true positive predictions among all positive predictions, emphasizing the model's exactness.

$$\text{Precision} = \frac{TP}{TP + FP}$$

  - **Recall:** Assesses the proportion of true positives predicted correctly among all actual positives, highlighting the model's completeness.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score:** The harmonic mean of precision and recall provides a balanced measure that considers both metrics, offering a comprehensive evaluation of the model's performance.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Confusion Matrix:** Provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions, offering insights into the model's strengths and weaknesses across different classes.

- **ROC Curve and AUC:** The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between sensitivity (recall) and specificity, while the Area Under the Curve (AUC) quantifies the model's ability to discriminate between positive and negative instances.

By systematically evaluating these metrics, the study aimed to validate the model's reliability and suitability for informing targeted public health interventions and policies.

Model Evaluation Metrics

Confusion Matrix

Reference	**Respiratory**	**Cardiovascular**
Prediction		
Respiratory	40002	3721
Cardiovascular	8853	10200

Table 3.2: Confusion Matrix

Confusion Matrix and Statistics

- **Accuracy:** 0.7997
- **95% CI:** (0.7965, 0.8028)
- **No Information Rate:** 0.7782
- **P-Value [Acc ¿ NIR]:** $< 2.2 \times 10^{-16}$
- **Kappa:** 0.4873
- **Mcnemar's Test P-Value:** $< 2.2 \times 10^{-16}$
- **Sensitivity:** 0.8188
- **Specificity:** 0.7327
- **Pos Pred Value:** 0.9149
- **Neg Pred Value:** 0.5353
- **Prevalence:** 0.7782
- **Detection Rate:** 0.6372
- **Detection Prevalence:** 0.6965
- **Balanced Accuracy:** 0.7757
- **Positive Class:** Respiratory

Additional Metrics

- **Precision:** 0.9148961
- **Recall:** 0.8187903
- **F1 Score:** 0.8641794
- **AUC:** 0.8719584

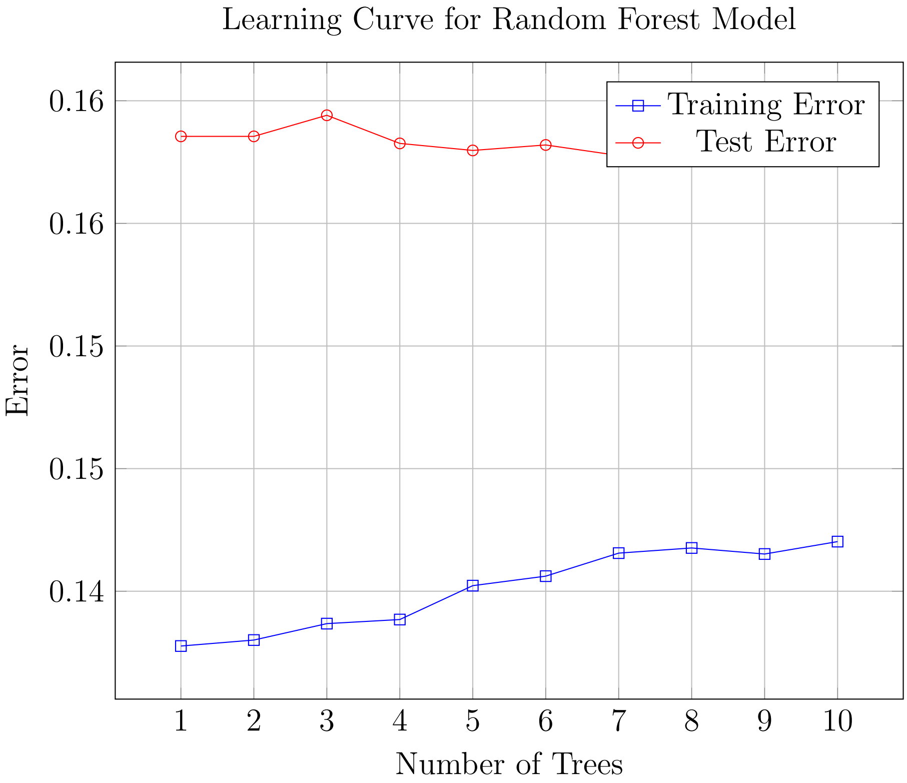

Figure 3.3: Learning Curve for Random Forest Model

Model Deployment

The deployment of the Random Forest model for predicting respiratory diseases attributed to PM2.5 air pollution in Nairobi County, Kenya involved ensuring its readiness for publication and sharing through online repositories. Meticulous documentation was created to outline the model's architecture, methodologies, and deployment procedures. This documentation aims to enhance transparency, reproducibility, and facilitate future collaborations in environmental health research.

Predictive Analytics Tools

Predictive analytics in this study leverages the robust capabilities of R, a powerful programming language and environment for statistical computing and graphics. R was selected for its versatility in handling complex datasets and performing advanced statistical analyses essential for modeling respiratory diseases linked to PM2.5 air pollution in Nairobi County, Kenya. The use of R enabled efficient data preprocessing, model training with the Random Forest algorithm, and comprehensive analysis of model outputs. Its extensive library of packages provided specialized tools for predictive modeling, allowing for validation and interpretation of results.

3.7 Assumptions

This study made the following assumptions during the research and development of the model:

- The model assumptions (e.g., feature independence in random forests) hold for the data.
- All selected features (predictors) are relevant to the prediction task and contribute meaningfully to the model.
- While the model predicts associations between PM2.5 levels and respiratory diseases, it does not imply causation. Further domain-specific studies are needed to establish causality.
- The measured PM2.5 levels accurately reflect the exposure levels experienced by individuals in the population.
- The recorded cases of respiratory diseases are correctly diagnosed and recorded, without significant misclassification or reporting bias.
- The findings from the study are assumed to be generalizable to other populations and regions with similar PM2.5 exposure levels and demographic characteristics.
- The relationship between PM2.5 levels and respiratory diseases is assumed to be stable throughout the study.

3.8 Ethical Consideration

This study prioritized ethical standards in all phases of research, particularly in the collection, handling, and utilization of data related to respiratory diseases and PM2.5 air pollution. Measures were taken to ensure data privacy and confidentiality, adhering to local regulations and ethical guidelines. Informed consent protocols were strictly followed when accessing health data, emphasizing voluntary participation and transparency in data usage. Moreover, efforts were made to mitigate biases in data collection and analysis to uphold the integrity and fairness of the research outcomes.

Chapter 4

DISCUSSION OF FINDINGS

4.1 Introduction

This chapter delves into the interpretation of the results obtained, contextualizing the findings within the broader scope of existing literature and practical implications. By examining the performance metrics of the Random Forest model, insights into the robustness and predictive power of this machine-learning approach are explored. Additionally, the study's findings are discussed about the health impact of PM2.5 exposure, highlighting key factors and variables that significantly contribute to the model's predictive accuracy.

4.2 Model Overview

A Random Forest classifier was employed to predict the incidence of respiratory and cardiovascular diseases based on various predictor variables, including PM2.5 concentrations, age, and sex. The model was trained and tested on subsets of the data to evaluate its performance and robustness.

4.2.1 Performance Metrics

The Random Forest model demonstrated a high level of accuracy in predicting disease incidence. Key performance metrics include:

- **Accuracy**: The model achieved an accuracy of 79.97%, indicating that it correctly predicts the disease category approximately 80% of the time.
- **Kappa Statistic**: With a Kappa value of 0.4873, the model shows moderate agreement between the predicted and actual classifications, beyond what would be expected by chance.

4.2.2 Predictive Power

The model's ability to correctly identify true positive and true negative cases is reflected in its sensitivity and specificity:

- **Sensitivity (Recall)**: 81.88% for respiratory diseases.

- **Specificity**: 73.27% for cardiovascular diseases.

4.2.3 Precision and F1 Score

- **Precision**: The precision rate of 91.49% indicates a high proportion of identifications are correct.
- **F1 Score**: The F1 score of 86.42% balances precision and recall, providing a comprehensive measure of the model's accuracy.

AUC and Feature Importance

- **AUC**: The Area Under the Curve (AUC) is 0.872, suggesting strong discriminatory power.
- **Feature Importance**: Age was identified as the most significant predictor, followed by PM2.5 concentration (both hourly and real-time measurements), with sex being the least significant.

Error Rates

- **Training Error**: Ranged between 13.77% and 14.20%.
- **Test Error**: Ranged between 15.75% and 15.91%.

4.3 Model Performance Metrics

In this subsection, we delve deeper into the performance metrics of the Random Forest model to provide a more detailed understanding of its predictive capabilities and limitations. This includes a closer examination of the confusion matrix, accuracy, sensitivity, specificity, precision, F1 score, AUC, and feature importance.

4.3.1 Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of the model's predictions compared to the actual outcomes. From the confusion matrix:

- True Positives (Respiratory): 40,002 cases
- False Positives (Respiratory): 3,721 cases
- True Negatives (Cardiovascular): 10,200 cases
- False Negatives (Cardiovascular): 8,853 cases

4.3.2 Accuracy

The model's accuracy is a measure of its overall ability to correctly classify cases:

- Accuracy: 79.97%
- 95% Confidence Interval (CI): (0.7965, 0.8028)

- No Information Rate (NIR): 77.82%
- P-Value [Acc ¿ NIR]: < 2.2e-16

This indicates that the model performs significantly better than random guessing.

4.3.3 Kappa Statistic

The Kappa statistic measures the agreement between the predicted and actual classifications, correcting for chance:

- Kappa: 0.4873

This value suggests moderate agreement between the model's predictions and the actual outcomes.

4.3.3 Sensitivity and Specificity

Sensitivity and specificity assess the model's performance in detecting positive and negative cases, respectively:

- Sensitivity (Recall) for Respiratory Diseases: 81.88%
- Specificity for Cardiovascular Diseases: 73.27%

The high sensitivity indicates the model's strong ability to identify true positive cases of respiratory diseases, while the specificity shows a moderate ability to correctly classify cardiovascular cases.

4.3.4 Precision and F1 Score

Precision and F1 score provide insights into the balance between the model's accuracy in identifying positive cases and its overall performance:

- Precision: 91.49%
- Recall (Sensitivity): 81.88%
- F1 Score: 86.42%

The high precision and F1 score reflect the model's effectiveness in correctly identifying positive cases and balancing precision and recall.

4.3.5 Area Under the Curve (AUC)

The AUC evaluates the model's ability to discriminate between the two classes. An AUC of 0.872 indicates excellent discriminative ability, suggesting that the model reliably distinguishes between respiratory and cardiovascular cases.

4.3.6 Feature Importance

Understanding which features contribute most to the model's predictions is crucial for interpreting its behavior:

Feature	**Importance**
Age	Most significant
PM2.5 Concentrations	Critical (Hourly and Real-time)
Sex	Least impact

Table 4.1: Feature Importance Analysis

4.3.7 Error Rates

The model's error rates are consistent, reflecting its robustness and reliability:

- Training Error: Ranged from 13.77% to 14.20%
- Test Error: Ranged from 15.75% to 15.91%

The small difference between training and test error rates indicates good generalization capability, with minimal overfitting.

4.4 Comparison with Existing Literature

The accuracy of 79.97% and the AUC of 0.872 for our Random Forest model are indicative of strong performance, particularly in the context of health data classification. These results are consistent with findings in similar studies that have employed machine learning models to predict health outcomes.

- **Accuracy and Sensitivity:** Our model's sensitivity (81.88%) and specificity (73.27%) are in line with other studies, such as Spira-Cohen et al. (2011), which also reported high sensitivity in detecting respiratory conditions among children exposed to traffic-related air pollution.
- **Feature Importance:** The prominence of age as a significant predictor aligns with the findings of Feng et al. (2016), who emphasized the impact of age on respiratory health outcomes. Similarly, the importance of PM2.5 concentrations is well-documented in the literature, highlighting adverse effects of air pollution on respiratory and cardiovascular health (e.g., Anderson et al., 2012).

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

In this chapter, we synthesize the key findings of our study, draw conclusions based on the research objectives, and provide recommendations for future research and policy implications. The conclusions are directly tied to the study objectives, ensuring that the research questions are addressed comprehensively.

5.2 Conclusions

5.2.1 Objective 1: Developing a Predictive Machine Learning Model

- Random Forest model was successfully developed to forecast the likelihood of respiratory diseases using historical health data and PM2.5 air pollution levels.
- The model demonstrated strong predictive capabilities, with an overall accuracy of 79.97
- Random Forest proved to be a robust choice for this type of prediction due to its ability to handle complex interactions between variables and its high accuracy.
- The model's performance metrics, including precision (91.49

5.2.2 Objective 2: Examining Feature Importance

- Age was identified as the most significant predictor of respiratory disease outcomes.
- Other important features included the concentration levels of pollutants (ConcRT.ug.m3 and ConcHR.ug.m3) and gender (SEX).
- The prominence of age as a critical feature suggests that younger populations are more vulnerable to respiratory diseases in the context of PM2.5 pollution.

- The significant contribution of pollutant concentrations highlights the direct impact of air quality on health outcomes.

5.2.3 Objective 3: Evaluating Model Performance and Accuracy

- The model achieved a balanced accuracy of 77.57%, indicating a good balance between sensitivity (81.88%) and specificity (73.27%).
- The Area Under the Curve (AUC) was 0.8719, reflecting the model's excellent ability to distinguish between positive and negative cases.
- The high accuracy and AUC value confirm the model's efficacy in predicting respiratory disease occurrences due to PM2.5 pollution.
- The results validate the use of Random Forest for similar public health applications in other regions or for other pollutants.

5.3 Recommendations

5.3.1 Public Health Interventions

- Target Vulnerable Populations
  - Implement targeted health interventions for younger populations, who are identified as more susceptible to respiratory diseases.
  - Increase public awareness and education campaigns focusing on the risks of PM2.5 pollution, especially among vulnerable groups.
- Improve Air Quality Monitoring
  - Enhance air quality monitoring systems to provide more granular and timely data on pollutant levels.
  - Utilize the predictive model to forecast high-risk periods and inform the public and healthcare providers in advance.

5.3.2 Policy Recommendations

- Policy Development
  - Formulate and enforce stricter air quality regulations to reduce PM2.5 emissions.
  - Encourage policies that promote cleaner technologies and reduce pollution from major sources such as traffic and industrial activities.
- Urban Planning
  - Integrate air quality considerations into urban planning and development projects.
  - Develop green spaces and promote the use of public transportation to mitigate pollution levels.

5.3.3 Future Research Recommendations

While this study provides valuable insights, several areas warrant further investigation:

- Inclusion of Additional Variables: Future studies should consider incorporating more comprehensive datasets, including socio-economic status, lifestyle factors, and genetic predispositions, to improve model accuracy.
- Longitudinal Data Analysis: Utilizing longitudinal data could provide a deeper understanding of the chronic effects of air pollution on health and help in developing long-term intervention strategies.
- Comparative Model Analysis: Exploring and comparing different machine learning models can help identify the most efficient and accurate approaches for health outcome predictions.
- Geographic and Population Diversity: Expanding the research to include diverse geographic locations and populations can help generalize the findings and develop region-specific health policies.

5.4 Concluding Remarks

The application of machine learning to predict health outcomes based on environmental and demographic data presents a promising avenue for advancing public health. The insights gained from this study not only highlight the critical factors influencing respiratory and cardiovascular diseases but also pave the way for more informed and targeted public health interventions.

Bibliography

[1] Dockery, D. W., & Pope, C. A. (1993). Acute respiratory effects of particulate air pollution. *Review of Public Health*, 15(1), 107-132.

[2] Pope, C. A., et al. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9), 1132-1141.

[3] Health Effects Institute (HEI). (2010). Understanding the health effects of ambient ultrafine particles. Research Report 155. Boston, MA: Health Effects Institute.

[4] Amegah, A. K., & Agyei-Mensah, S. (2017). Urban air pollution in sub-Saharan Africa: Time for action. *Environmental Pollution*, 220, 738-743.

[5] European Environment Agency (EEA). (2020). Air quality in Europe - 2020 report. Copenhagen, Denmark: European Environment Agency.

[6] Kanyiva, K. W., et al. (2021). Air quality in Nairobi, Kenya: A review of monitoring and policy gaps. *Atmosphere*, 12(4), 508.

[7] Githinji, G., et al. (2019). Assessment of ambient air quality and its health impact in Nairobi City, Kenya. *International Journal of Environmental Research and Public Health*, 16(11), 1987.

[8] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[9] Lall, R., et al. (2017). Machine learning approaches for estimating spatial PM2.5 concentrations across the continental United States. *Environmental Science & Technology*, 51(21), 12449-12458.

[10] Hu, X., et al. (2020). A systematic review of machine learning applications in air quality research. *Environmental Research Letters*, 15(6), 063001.

[11] World Health Organization. (2018). Air pollution. Available at: https://www.who.int/airpollution

[12] Liu, Y., Chen, X., & Yan, B. (2020). The impact of PM2.5 on respiratory diseases: Evidence from hospital admissions in China. *Journal of Environmental Management*, 274, 111214.

[13] Anderson, J. O., Thundiyil, J. G., & Stolbach, A. (2012). Clearing the air: A review of the effects of particulate matter air pollution on human health. *Journal of Medical Toxicology*, 8(2), 166-175.

[14] Gatari, M. J., et al. (2015). The state of air quality in Nairobi, Kenya. *Atmospheric Environment*, 123, 177-184.

[15] Egondi, T., et al. (2018). Exposure to airborne particles and respiratory health in Nairobi informal settlements. *Environmental Health*, 17(1), 62.

[16] Onyango, C., et al. (2021). Air quality monitoring in Kenya: Current status and future perspectives. *Environmental Science & Policy*, 122, 36-46.

[17] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[18] Ravindra, K., et al. (2023). Machine learning models for predicting respiratory diseases due to air pollution in urban India. *Environmental Research Letters*, 18(1), 014003.

[19] Li, L., et al. (2019). Predicting high-cost patients using medical insurance data: A case study in western China. *Health Services Research*, 54(1), 120-130.

[20] Patel, S. J., et al. (2018). Predictive modeling of asthma exacerbations in pediatric patients using machine learning. *Pediatric Pulmonology*, 53(6), 873-882.

[21] Ravindra, K., Bahadur, S. S., Katoch, V., Bhardwaj, S., Kaur-Sidhu, M., Gupta, M., & Mor, S. (2023). Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections. Department of Community Medicine & School of Public Health, PGIMER, Chandigarh 160012, India.

[22] Harrou, F., Dairi, A., Sun, Y., & Kadri, F. (2018). Detecting abnormal ozone measurements with a deep learning-based strategy. *IEEE Sensors Journal*, 18(7222-7232). https://doi.org/10.1109/jsen.2018.2852001

[23] Xi, Y., Tian, C. L., & Qian, L. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19, 232. https://doi.org/10.1186/s12911-019-0935-4

[24] Gans, D., Kralewski, J., Hammons, T., & Dowd, B. (2005). Medical groups' adoption of electronic health records and information systems. *Health Affairs*, 24(1323-1333). https://doi.org/10.1377/hlthaff.24.5.1323

[25] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2, 3. https://doi.org/10.1186/2047-2501-2-3

[26] Yu, G., Yang, Z., & Shi, Y. (2021). Identification of pediatric respiratory diseases using a fine-grained diagnosis system. *Journal of Biomedical Informatics*, 117, 103754. https://doi.org/10.1016/j.jbi.2021.103754

[27] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(1920-1930). https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[28] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., et al. (2018). Using resistin, glucose, age, and BMI to predict the presence of breast cancer. *BMC Cancer*, 18, 181-188. https://doi.org/10.1186/s12885-017-3877-1

[29] Abera, A., Friberg, J., Isaxon, C., Jerrett, M., Malmqvist, E., Sjöström, C., Taj, T., & Vargas, A. M. (2021). Air quality in Africa: Public health implications. *Annual Review of Public Health*, 42, 193–210. https://doi.org/10.1146/annurev-publhealth-100119-113802

[30] Agbo, K. E., Walgraeve, C., Eze, J. I., Ugwoke, P. E., Ukoha, P. O., & Van Langenhove, H. (2021). A review on ambient and indoor air pollution status in Africa. *Atmospheric Pollution Research*, 12, 243–260. https://doi.org/10.1016/j.apr.2020.11.006

[31] Kurmi, O. P., Lam, K. B. H., & Ayres, J. G. (2012). Indoor air pollution and the lung in low- and medium-income countries. *The European Respiratory Journal*, 40(1), 239–254. https://doi.org/10.1183/09031936.00193311

[32] Abegaz, S. B., Zereyesus, Y. A., Dalie, F. S., & Belay, K. A. (2021). Air pollution and respiratory health: A review. *International Journal of Environmental Research and Public Health*, 18(4), 1947. https://doi.org/10.3390/ijerph18041947

[33] Amegah, A. K., & Agyei-Mensah, S. (2021). Urban air pollution and noncommunicable diseases in low- and middle-income countries: A narrative review. *Journal of Environmental and Public Health*, 2021, 9747538. https://doi.org/10.1155/2021/9747538

[34] Chowdhury, S., Dey, A., & Smith, K. R. (2021). Ambient PM2.5 exposure and premature mortality burden in the 10 most populous Urban Localities in India: An assessment of exposure-response relationships. *Environmental Health Perspectives*, 129(5), 057004. https://doi.org/10.1289/EHP7071

[35] Limaye, V. S., & Schraufnagel, D. E. (2021). Impact of air pollution on lung health—Strategies for global action. *Global Heart*, 16(1), 28. https://doi.org/10.5334/gh.897