Digital Transformation To Reduce The Air Pollution Based On IoT and Machine Learning Approaches

June 11, 2023

Student name: Mahmoud Rumaneh Student number: 20120103

https://www.overleaf.com/read/wngdrxcpfcdx

Centre name: Al-Hussein Technical University

Tutor: Dr. Eyad Taqieddin, Eng. Qusai Ismail

Unit: Computing research project

Abstract

Air pollution is a major problem that affects people all over the world. It can cause a variety of health problems, including respiratory problems, heart disease, and cancer. Air pollution can also damage the environment which leads to climate change and other problems.in this research paper explores using IoT and machine learning to tackle air pollution in Africa. IoT devices collect data on air pollution, revealing sources and patterns. Machine learning models analyze the data and make predictions, aiding in the development of effective strategies for reducing pollution. The study emphasizes the importance of real-time monitoring and proactive interventions. The Random Forest Regression model performs well in predicting pollution levels. Overall, the research highlights the potential of digital transformation through IoT and machine learning to address air pollution challenges and provides a foundation for future work in this area.

1 Section One: Introduction to Environmental Impact of Digital Transformation and Introduction to Subject

The history of air pollution is long and complex, with complaints about its effects on human health and the environment dating back to ancient Athens and Rome. However, air quality worsened considerably during the Industrial Revolution, when coal became widespread as a fuel source. Early laws to control air pollution were generally weak and ineffective due to the importance of coal-fueled steam power to economic growth. It was not until the mid-twentieth century, after major air pollution episodes such as London's 'Great Smog', that stringent national laws to abate smoke were introduced to clear the skies over industrial cities. Today, smoke pollution is still a significant environmental problem in many industrial cities in developing countries[1]. After the Second World War, new threats such as acid rain, photochemical smog, ozone depletion, and climate change emerged, requiring the cooperation of scientific experts across borders and collective political action. The success of Montreal Protocol stands as a successful example of international environmental governance, but reducing greenhouse gas emissions will require a strong commitment to international cooperation, particularly as global warming is still a difficult concept for many to grasp [1].

Air pollution is one of the most significant threats to public and individual health, not just because of its impact on climate change but also due to increased morbidity and mortality. Particulate Matter (PM), which comprises particles of very small diameter, is one of the major pollutants responsible for respiratory and cardiovascular diseases, reproductive and central nervous system dysfunctions, and cancer. Ground-level ozone, which is harmful to humans when present in high concentrations, also affects the respiratory and cardiovascular systems. Nitrogen oxide, sulfur dioxide, Volatile Organic Compounds (VOCs), dioxins, and polycyclic aromatic hydrocarbons (PAHs) are other air pollutants that are harmful to human health. Carbon monoxide can lead to direct poisoning when breathed in at high levels, while heavy metals such as lead can lead to direct poisoning or chronic intoxication depending on exposure. Respiratory problems such as Chronic Obstructive Pulmonary Disease (COPD), asthma, bronchiolitis, lung cancer, cardiovascular events, central nervous system dysfunctions, and cutaneous diseases are the primary diseases that occur due to these substances. It's important to know how environmental pollution and climate change can impact the geographical distribution of infectious diseases and natural disasters. The only way to tackle this problem is through public awareness and a multidisciplinary approach by scientific experts. National and international organizations must address the emergence of this threat and propose sustainable solutions to reduce the impact of air pollution on public and individual health [2].

As well as, digital transformation has an impact on organizations and society, particularly in the areas of environmental sustainability. Digital transformation refers to the use of new digital technologies to improve business operations and markets, creating opportunities for fundamental changes in institutions. This process has been fueled by technologies such as the Internet of Things, big data analysis, cloud computing, mobile technologies, and artificial intelligence. Digital transformation has wide-ranging effects beyond consumer behavior and organizational improvements, including impacts on healthcare and social dynamics. Additionally, digital transformation is expected to affect environmental sustainability, as it enables green technology innovation, accelerates human capital accumulation, increases environmental information disclosure, and strengthens environmental governance [3].

The utilization of digitalization is crucial for resource-based industries to overcome the limitations of resources and environmental constraints. By a paper that analyzed the direct impact and mechanisms of digital transformation on the environmental performance of resource-based companies in China, using data from the country's A-share resource-based listed enterprises. The research indicates that digital transformation plays a significant role in boosting the environmental performance of companies by promoting green technology innovation, enhancing human capital accumulation, increasing environmental information disclosure, and strengthening environmental governance. Moreover, the impact of digital transformation on the environmental performance of resource-based enterprises is more pronounced for state-owned, large-scale, and high-tech companies. Additionally, the study highlights that the eastern region and areas with more stringent environmental regulations exhibit a more significant impact of digital transformation on the environmental performance of resource-based enterprises. The research provides practical recommendations for resource-based industries to leverage digital dividends for achieving high-quality development [4].

This paper provides an explanation of the role and importance of the Internet of Things and Machine Learning in preserving the environment, how data related to air pollution was collected through it, how air pollution affects people's lives and the dangers of that, and how people are unaware of air pollutants and their impact on their health, and how we can come up with a beneficial result to reduce Air pollution errors and its treatment also through the Internet of Things and building machine learning.

The collected data can be used to analyze the sources of air pollution and develop strategies to mitigate its harmful effects. In this paper, machine learning approaches will be employed to reduce air pollution. By leveraging the power of machine learning algorithms, large amounts of data can be processed and analyzed to identify patterns and trends. These insights can then be used to predict future air pollution levels and inform the development of targeted interventions aimed at improving air quality in specific areas [5].

Machine learning techniques can help in optimizing the operation of IoT devices and sensors, enabling them to effectively monitor air pollution levels and identify potential sources. Additionally, by incorporating machine learning algorithms into air pollution control systems, real-time data can be continuously analyzed, allowing for prompt responses and adjustments to minimize pollution levels. Ultimately, the integration of machine learning and IoT technologies holds great potential in combating air pollution and creating healthier environments for communities worldwide [5].

2 Section Two: Title, objective, research questions

Title: Digital Transformation To Reduce The Air Pollution Based On IoT and Machine Learning Approaches.

Objective: The goal is to show how the IoT is useful for collecting data from air pollution and analyzing it by machine learning approaches to get results that help to find a way to reduce air pollution.

Questions: How can we reduce the air pollution?

How can the air pollution affect human life?

How can the IoT help to reduce and measure air pollution through the collection of data?

How to use the data collected using IoT to train a machine learning model to reduce air pollution?

3 Section Three: Reasons for choosing this research project

1. Relevance and Timeliness: Air pollution is a pressing global issue that affects the health and well-being of populations worldwide. By addressing this topic, my research project

aligns with current environmental concerns and contributes to ongoing efforts to mitigate air pollution.

2. Innovation and Technological Advancements: Incorporating machine learning approaches and IoT technologies in my research project demonstrates a forward-thinking and innovative approach. It showcases the potential of emerging technologies to tackle complex environmental challenges and offers new possibilities for effective pollution reduction strategies.
3. Practical Application: The use of IoT for data collection, as demonstrated by the authors of the Urban Air Pollution Challenge, offers a practical and scalable solution for monitoring air quality. By analyzing this dataset, I can gain insights into the factors influencing air pollution and develop machine learning models that can contribute to more accurate predictions and informed decision-making in pollution reduction efforts.
4. Interdisciplinary Nature: My research project bridges the domains of environmental science, data analysis, machine learning, and IoT. By combining these fields, I have the opportunity to provide a holistic perspective on the challenges of air pollution and showcase the potential of interdisciplinary approaches in addressing complex environmental problems.
5. Contribution to Digital Transformation: Digital transformation, including the use of IoT and machine learning, has the potential to revolutionize environmental management and contribute to sustainable development. By exploring the impact of digital transformation on air pollution reduction, my research project sheds light on the broader implications of technological advancements in environmental conservation.
6. Societal Impact: The outcome of my research project has the potential to make a positive impact on society by providing insights and solutions for reducing air pollution. By leveraging machine learning techniques, I can contribute to the development of more effective pollution reduction strategies, potentially leading to improved air quality, better public health outcomes, and a more sustainable environment.

---

4 Section Four: Literature sources searched

This paper [6] shows how the industrial boom has led to an increase in the number of industries and factories, which have contributed significantly to environmental pollution. With the rapid advancement of technology, the need for continuous monitoring of air quality has become more important than ever. However, most of the existing systems are expensive and complicated, resulting in delays in implementing effective measures to reduce air pollution. In recent years, there has been a growing interest in developing low-cost air quality monitoring systems to address this issue. These systems are designed to be affordable and accessible, making it possible for individuals and communities to monitor air quality in real-time. However, dilatory increases in low-cost air quality monitoring systems have been responsible for overall degradation. This is because these systems may not be reliable or accurate, leading to incorrect data and ineffective measures to mitigate pollution. To address these concerns, researchers have proposed various low-cost air quality monitoring systems that are simple in design, mobile, and affordable. These systems typically consist of gas sensors, a communication module, a cloud server, and a mobile application. For example, a recent study proposed an air pollution detection and monitoring system that incorporates various gas sensors, a GSM module, a cloud server, and a mobile application. The system allows users to easily access air quality data from the server and app and also features an alert system that notifies responsible officials when pollution parameters exceed standard permissible limits. Furthermore, the overall cost of implementing this proposed system was BDT 3500 (USD 40), which is lower than previously implemented systems in Bangladesh. Such systems offer a promising solution for low-cost air quality monitoring and management, making it possible to mitigate environmental pollution more effectively. In conclusion, low-cost air quality monitoring systems offer a practical and affordable solution to the problem of environmental pollution caused by industries and factories. The proposed air pollution detection and monitoring system presented in this study provides a promising example of how such systems can be developed to be reliable, accurate, and affordable.

This paper [7] shows how the issue of urban air pollution has reached an alarming state across India, with most cities facing poor air quality that fails to meet the standards for good health. To address this problem, there is a need to develop an air pollution measurement and prediction system that can be used in smart cities. In recent years, researchers have proposed various air pollution measurement and prediction systems that use pollution detection sensors and cloud services to acquire, process, and analyze air quality data. These systems typically involve low-cost embedded boards and gas sensors for data acquisition, cloud services for data storage and analysis, and machine learning algorithms for pollution prediction. For example, a recent study proposed a system that acquires carbon dioxide and carbon monoxide levels in the air, along with GPS location, using pollution detection sensors and uploads them into Microsoft Azure cloud services. The system uses a low-cost embedded Beagle bone board and gas sensors for data acquisition and an Azure machine learning service for pollution prediction based on previous data. The data is then represented by the Power BI tool for further analysis. The proposed system has been implemented and found to be useful for monitoring and reducing pollution in smart cities by identifying and avoiding pollution causes. The calibrated gas sensor data is fetched from the sensors and successfully uploaded into the cloud, where it is utilized by different cloud services to make the data meaningful. Such air pollution measurement and prediction systems offer a practical solution to the problem of poor air quality in smart cities, providing real-time data and insights for effective pollution management. In conclusion, the proposed system offers a promising example of how such systems can be developed to be affordable, reliable, and effective in mitigating the negative effects of urban air pollution on public health.

In this paper [8], one recent study proposed a three-phase air pollution monitoring system that uses an IoT kit comprising gas sensors, an Arduino integrated development environment (IDE), and a Wi-Fi module. The IoT kit can be physically placed in various cities to monitor air pollution, with the gas sensors gathering data from the air and forwarding it to the Arduino IDE. The IDE then transmits the data to the cloud via the Wi-Fi module. To enable users to access relevant air quality data from the cloud, the study also developed an Android application called IoT-Mobair. The application can predict the pollution level of the entire route if a user is traveling to a destination, and display a warning if the pollution level is too high. The proposed system is analogous to Google Traffic or the navigation application of Google Maps. Additionally, the air quality data can be used to predict future air quality index (AQI) levels. Such IoT-based air pollution monitoring systems offer a practical solution to the problem of global air pollution, providing real-time data and insights for effective pollution management. Furthermore, the proposed system is low-cost, mobile, and easy to use, making it a promising example of how IoT technologies can be harnessed for environmental monitoring and protection. As well as, the proposed system presents a significant improvement over existing air pollution monitoring systems and could help to address the challenges posed by air pollution and its negative impact on public health.

The authors of this paper [9] conducted a systematic mapping study using a five-step methodology to identify and analyze the research status of IoT-based air pollution monitoring systems for smart cities. The study reviewed 55 proposals, some of which had been implemented in a real environment. The authors analyzed and compared these proposals based on various parameters defined in the mapping study and identified several challenges to air quality monitoring systems' implementation in the smart city context. The study revealed that several IoT-based air pollution monitoring systems have been proposed to measure different pollutants using interconnected sensors. These systems have the potential to provide accurate and timely information on air quality levels in smart cities. The authors classified the proposals into four categories based on their approach: monitoring and prediction, monitoring and control, monitoring and feedback, and monitoring and analysis. The authors highlighted several challenges for the implementation of air quality monitoring systems in smart cities. One challenge is the high cost associated with the installation and maintenance of sensors, which could limit the scalability of such systems. Additionally, the lack of standards for data acquisition and processing, data privacy and security concerns, and the need for interoperability between different systems pose challenges to the implementation of these systems.

This paper [10] shows how air pollution is a major problem in urban areas, with gas emissions from cars being the most significant contributor. These emissions contain various pollutant gases, including carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), particulate matter (PM), and Sulphur dioxide (SO2). Monitoring air quality in real-time using IoT devices has emerged as a potential solution to this problem. This paper aims to explore the use of IoT-based air quality monitoring systems and their effectiveness in addressing the issue of air pollution in urban areas. The Environmental Protection Agency (EPA) has established guidelines for measuring pollutant gases in the air, including CO, NO2, O3, PM, and SO2. These guidelines use several methods to calculate the concentration of these chemicals in the air. One promising approach to monitoring air quality in real-time is using IoT devices equipped with a set of sensors that measure air quality at the street level. This paper aims to investigate the relationship between traffic volume and the Air Quality Index (AQI), as defined by EPA guidelines. The authors used Multiple Linear Regression (MLR) to create a mathematical model for this relationship. The model was tested on a street in the city of Melbourne, Florida, and the results were analyzed to understand the impact of traffic volume on air quality.

The authors in this paper [11] stated that the use of big datasets in air pollution epidemiology presents both potential and challenges. In response, researchers have explored alternative methods such as data mining and machine learning algorithms for making predictions, finding patterns, and extracting information. The authors conducted a systematic search of research articles and identified 47 that applied data mining and machine learning methods in air pollution epidemiology. The research articles were grouped into three areas of interest: source apportionment, forecasting/prediction of air pollution/quality or exposure, and generating hypotheses. The authors noted that early applications had a preference for artificial neural networks, while more recent work has applied decision trees, support vector machines, k-means clustering, and the APRIORI algorithm. The majority of the research has been conducted in Europe, China, and the USA. The authors identified two emerging areas of data mining with good potential for future applications in air pollution epidemiology: deep learning and geospatial pattern mining. They concluded that data mining methods have been used to address a range of issues in air pollution epidemiology and that the potential for supporting this field continues to grow with advancements in data mining techniques related to temporal and geospatial mining and deep learning.

Paper Title	Main objective
Design and Implementation of an IoT-Based Air Pollution Detection and Monitoring System [6].	The authors in this paper showed how the industrial boom has led to an increase in the number of industries and factories, which have contributed significantly to environmental pollution.
IoT-based air pollution monitoring and predictor system on Beagle bone black [7].	The authors in this paper showed how the issue of urban air pollution has reached an alarming state across India, with most cities facing poor air quality that fails to meet the standards for good health.
Internet of Things Mobile–Air Pollution Monitoring System (IoT-Mobair) [8].	The authors in this paper showed one recent study proposed a three-phase air pollution monitoring system that uses an IoT kit comprising gas sensors, an Arduino integrated development environment (IDE), and a Wi-Fi module.
IoT-based air quality monitoring systems for smart cities: A systematic mapping study [9].	The authors in this paper conducted a systematic mapping study using a five-step methodology to identify and analyze the research status of IoT-based air pollution monitoring systems for smart cities. showed
IoT Based: Air Quality Index and Traffic Volume Correlation [10].	The authors in this paper showed how air pollution is a major problem in urban areas, with gas emissions from cars being the most significant contributor. These emissions contain various pollutant gases, including carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), particulate matter (PM), and Sulphur dioxide (SO2).
A systematic review of data mining and machine learning for air pollution epidemiology [11].	The authors in this paper stated that the use of big datasets in air pollution epidemiology presents both potential and challenges. In response, researchers have explored alternative methods such as data mining and machine learning algorithms for making predictions, finding patterns, and extracting information.

5 Section Five: Activities and timescales

5.1 First Semester

Task	Start Date	Finish Date	Duration
Choose the general research topic, Write an introduction to the theme, and submit the logbook.	25/11/2022	1/12/2022	Week
More research about the chosen topic, and modify the introduction, and submit the logbook.	2/12/2022	8/12/2022	Week
Summarize the most relevant paper, and submit the 5 most relevant papers to their research topic, and submit the logbook.	9/12/2022	15/12/2022	Week
Finalize the research idea and state clearly the research questions, summarize three to four papers from the chosen papers, as part of the literature review, and submit the logbook.	16/12/2022	22/12/2022	Week
Summarize the rest of the chosen papers, as part of the literature review, Suggest a possible source of data, and submit the logbook.	23/12/2022	29/12/2022	Week
Review the research questions, review the research source of data, and submit the logbook.	30/12/2022	5/1/2023	Week
Suggest the research approach and methodologies, and submit the logbook.	6/1/2023	12/1/2023	Week
Continue working on the proposal, and submit the logbook.	13/1/2023	19/1/2023	Week

5.2 Second Semester

Task	Start Date	Finish Date	Duration
Searching for data-set about using IoT for air pollution	2/3/2023	30/3/2023	28 Days
Making data cleaning	31/3/2023	6/4/2023	Week
Refining the collected data and making dashboards	7/4/2023	21/4/2023	2 Weeks
Modeling statistics and building machine learning	22/4/2023	6/5/2023	2 Weeks
Conclusion	7/5/2023	21/5/2023	2 Weeks

6 Section Six: Dataset

6.1 Data Collection

The data for the Urban Air Pollution Challenge [12] was collected using a specific methodology outlined by the challenge organizers for Africa. The objective of the challenge was to predict the daily concentration of PM2.5 particulate matter, which is a significant air pollutant associated with health concerns when levels in the air are high. To gather the necessary data, the authors employed three main sources.

Firstly, ground-based air quality sensors equipped with Internet of Things (IoT) technology were utilized. These sensors measured the target variable, which is the concentration of PM2.5 particles in the air. The data provided by these sensors included the daily mean concentration of PM2.5, as well as additional information such as the minimum and maximum readings for the day, the variance of the readings, and the total count of sensor readings used to compute the target value. It is important to note that this data was only available for the training set, and the participants of the challenge were tasked with predicting the target variable for the test set [12].

The second source of data was the Global Forecast System (GFS) for weather data. The GFS data provided meteorological information such as humidity, temperature, and wind speed. These weather variables were considered potential inputs for the predictive models, as they have a known influence on the concentration of PM2.5 particles in the air [12].

The third source of data was the Sentinel 5P satellite, which was used to monitor various pollutants in the atmosphere. The authors queried the offline Level 3 (L3) datasets available in Google Earth Engine for each pollutant of interest. For example, they retrieved relevant data from the Sentinel 5P dataset for NO2 (nitrogen dioxide). The data included key measurements such as NO2_column_number_density, representing the concentration of NO2, as well as metadata like the satellite altitude. The challenge organizers recommended focusing on the key measurements, particularly the column_number_density or the tropospheric_X_column_number_density, which provide measurements closer to the Earth's surface [12].

This dataset contains 82 features. Some locations might have lacked sensor readings for certain days, leading to the exclusion of those specific rows. Additionally, data gaps were present, particularly in the satellite data for CH4 (methane) [12].

In conclusion, the authors of the Urban Air Pollution Challenge collected data from ground-based air quality sensors utilizing IoT technology, the Global Forecast System (GFS) for weather data, and the Sentinel 5P satellite for monitoring various pollutants. This diverse range of data sources provided crucial information for participants to predict the PM2.5 particulate matter concentration in the air, addressing the challenge's objective [12].

You can see all the dataset's data from here: [13]

6.2 Data Analysis and Visualization

The figure 1 demonstrates the relationship between the precipitable water in the entire atmosphere and the air pollution level. The x-axis represents the precipitable water, and the y-axis represents the air pollution level. Precipitable water refers to the total amount of water vapor present in a vertical column of the atmosphere. While water vapor can form clouds and precipitation, the scatter plot does not exhibit a distinct relationship between precipitable water and PM2.5 concentrations. Other factors, such as pollutant emissions and atmospheric conditions, may significantly impact air pollution levels. So, based on the scatter plot, there is no clear correlation between precipitable water and PM2.5 concentrations. Therefore, it is less likely that precipitable water alone has a significant influence on air pollution levels.

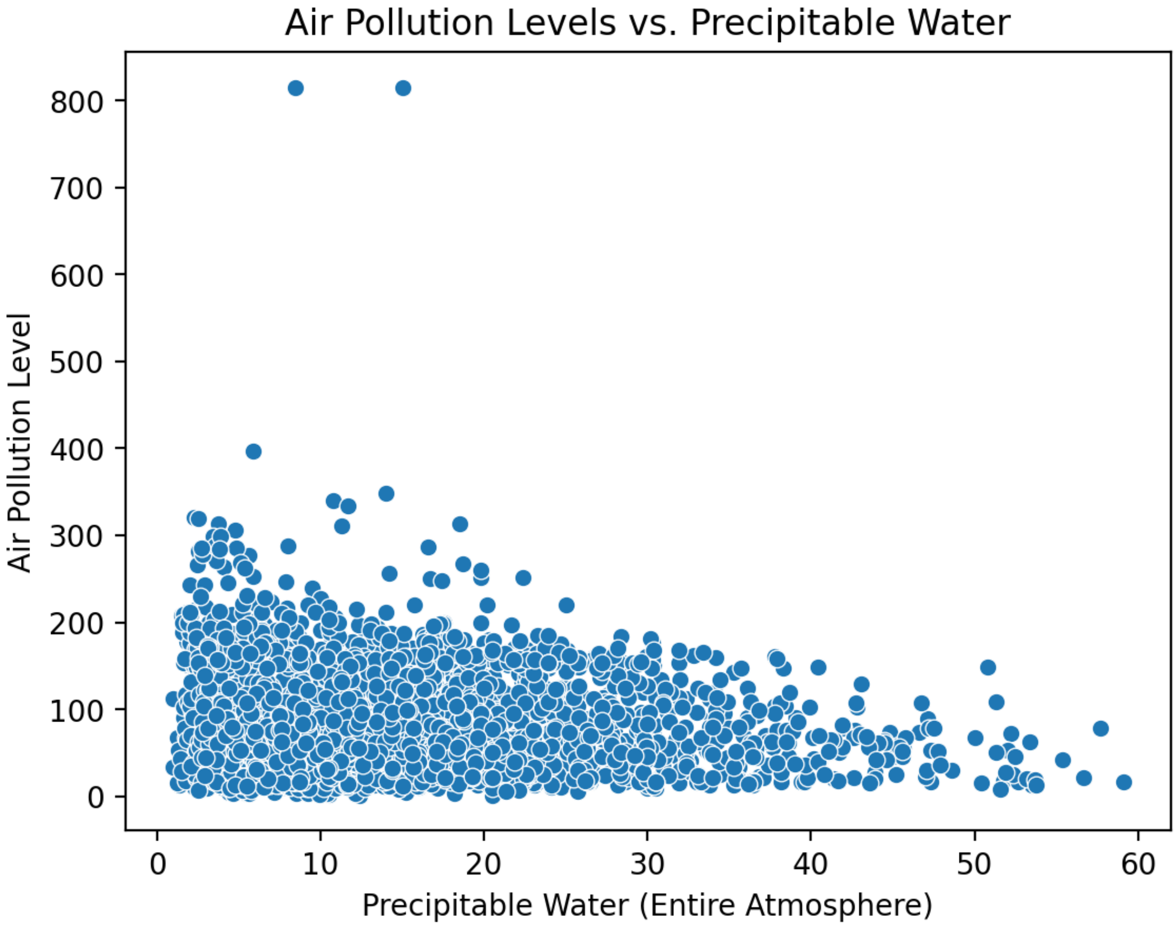

Figure 1: Air Pollution Levels vs. Precipitable Water

The figure 2 displays the relationship between the relative humidity at 2m above the ground and the air pollution level. The x-axis represents the relative humidity, and the y-axis represents the air pollution level.

This feature seems to have a positive correlation with the air pollution level. The higher relative humidity is likely to contribute to increased PM2.5 concentrations. As relative humidity increases, it indicates a higher moisture content in the air. This can result in the accumulation of particulate matter and pollutants, including PM2.5, as moisture can act as a carrier for these particles. High humidity levels can also contribute to the condensation of pollutants, making them more likely to remain in the atmosphere and increase air pollution levels.

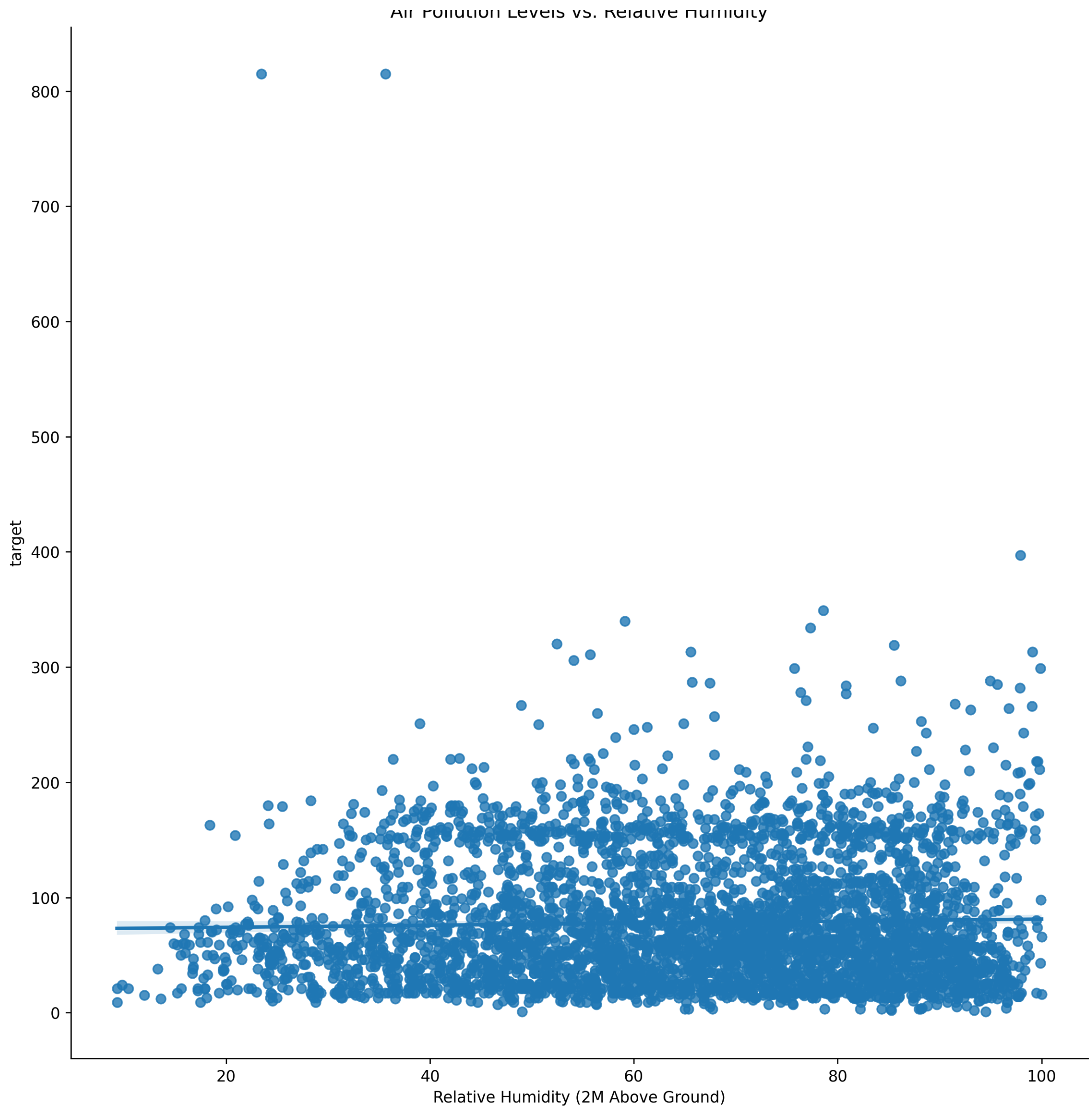

Figure 2: Air Pollution Levels vs. Relative Humidity

The figure 3 depicts the relationship between the NO2 slant column number density and the air pollution level. The x-axis represents the NO2 slant column number density, and the y-axis represents the air pollution level.

NO2 is a pollutant primarily emitted from combustion processes, such as vehicle emissions and industrial activities. The slant column number density represents the vertical column of NO2 in the atmosphere. Higher values indicate higher concentrations of NO2, which can react with other pollutants and contribute to the formation of PM2.5. This feature seems to have a positive correlation with the air pollution level. Higher NO2 slant column number density is likely to contribute to increased PM2.5 concentrations.

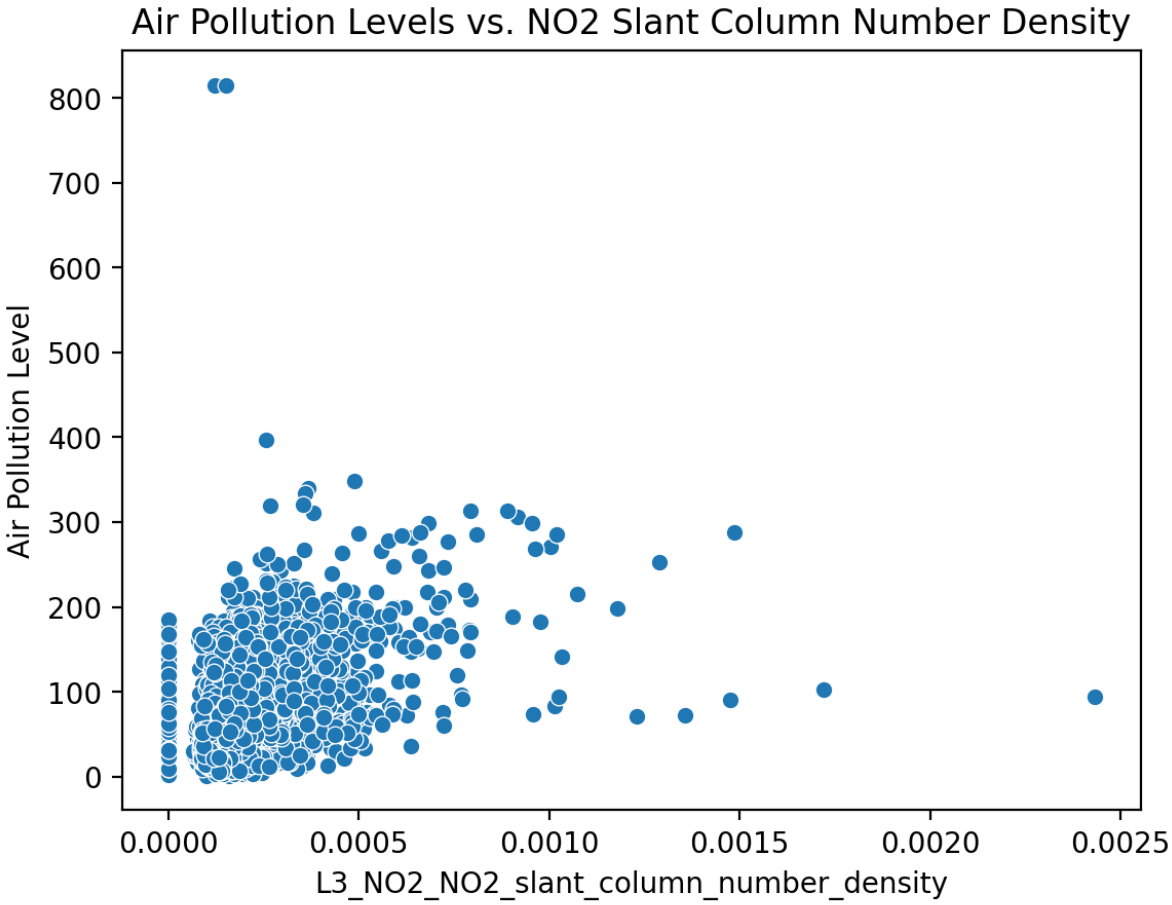

Figure 3: Air Pollution Levels vs. NO2 Slant Column Number Density

The figure 4 depicts the relationship between the CO column number density and the air pollution level. The x-axis represents the CO column number density, and the y-axis represents the air pollution level.

The scatter plot shows a positive correlation between CO column number density and PM2.5 levels. Higher CO levels indicate increased emissions from combustion processes, such as vehicle exhaust and industrial activities. These emissions contribute to air pollution and can lead to higher PM2.5 concentrations. Based on the positive correlation, it can be concluded that higher CO column number density is likely to contribute to increased PM2.5 concentrations. Efforts to reduce CO emissions can help mitigate air pollution and potentially lower PM2.5 levels.

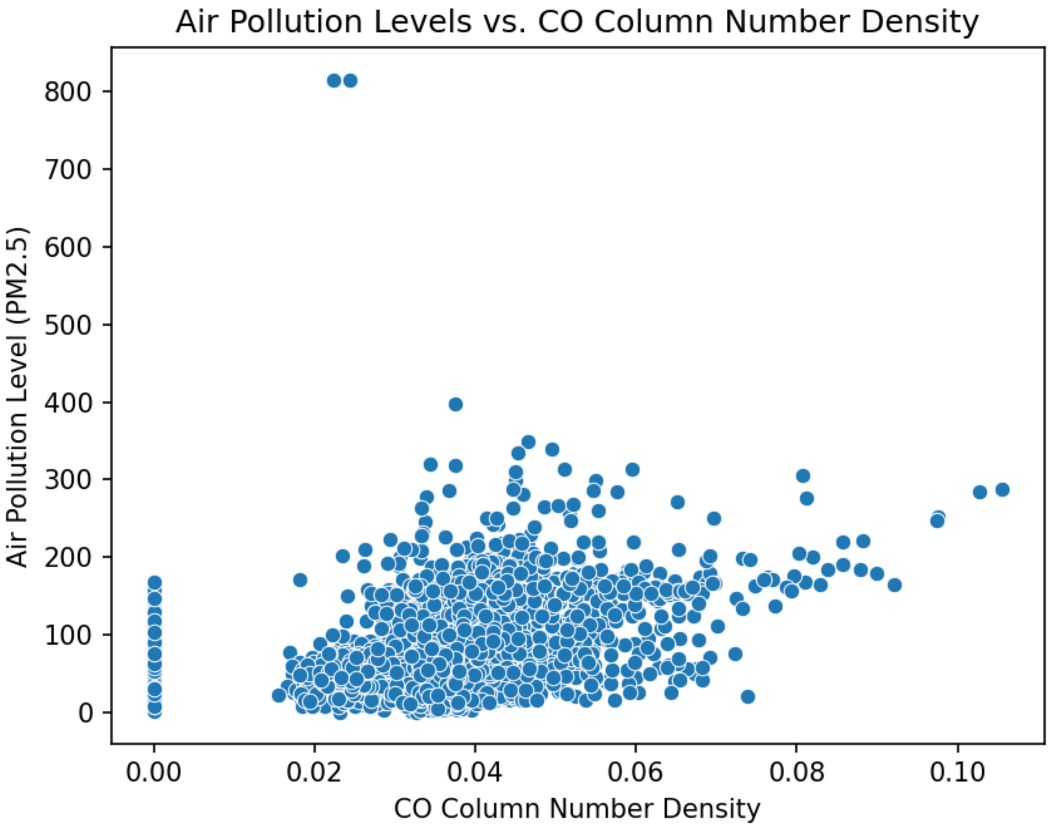

Figure 4: Air Pollution Levels vs. CO Column Number Density

The figure 5 depicts the relationship between the HCHO (formaldehyde) slant column number density and the air pollution level. The x-axis represents the HCHO (formaldehyde) slant column number density, and the y-axis represents the air pollution level.

The scatter plot reveals a positive correlation between HCHO slant column number density and PM2.5 levels. Higher HCHO levels are associated with emissions from industrial sources and vehicle exhaust, which are known contributors to air pollution. As HCHO concentrations increase, it is likely that PM2.5 concentrations will also rise. Based on the positive correlation, it can be concluded that higher HCHO slant column number density is likely to contribute to increased PM2.5 concentrations. Strategies aimed at reducing HCHO emissions can help mitigate air pollution and potentially lower PM2.5 levels.

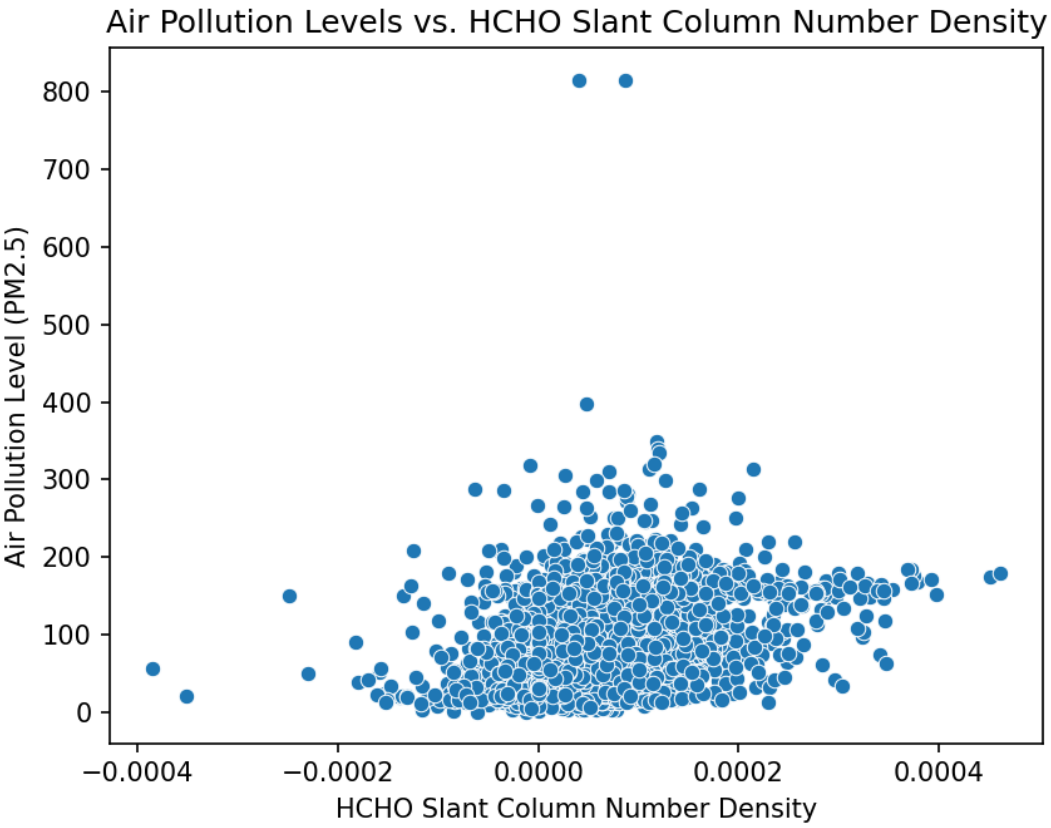

Figure 5: Air Pollution Levels vs. HCHO Slant Column Number Density

The figure 6 depicts the relationship between the absorbing aerosol index (AAI) and the air pollution level. The x-axis represents the absorbing aerosol index (AAI), and the y-axis represents the air pollution level.

The scatter plot shows a positive correlation between the AAI and PM2.5 levels. Higher AAI values indicate the presence of absorbing aerosols, which can originate from sources like industrial emissions and biomass burning. These aerosols contribute to air pollution and can lead to increased PM2.5 concentrations. Decision: Based on the positive correlation, it can be concluded that a higher absorbing aerosol index is likely to contribute to increased PM2.5 concentrations. Implementing measures to reduce the emissions of absorbing aerosols can help mitigate air pollution and potentially lower PM2.5 levels.

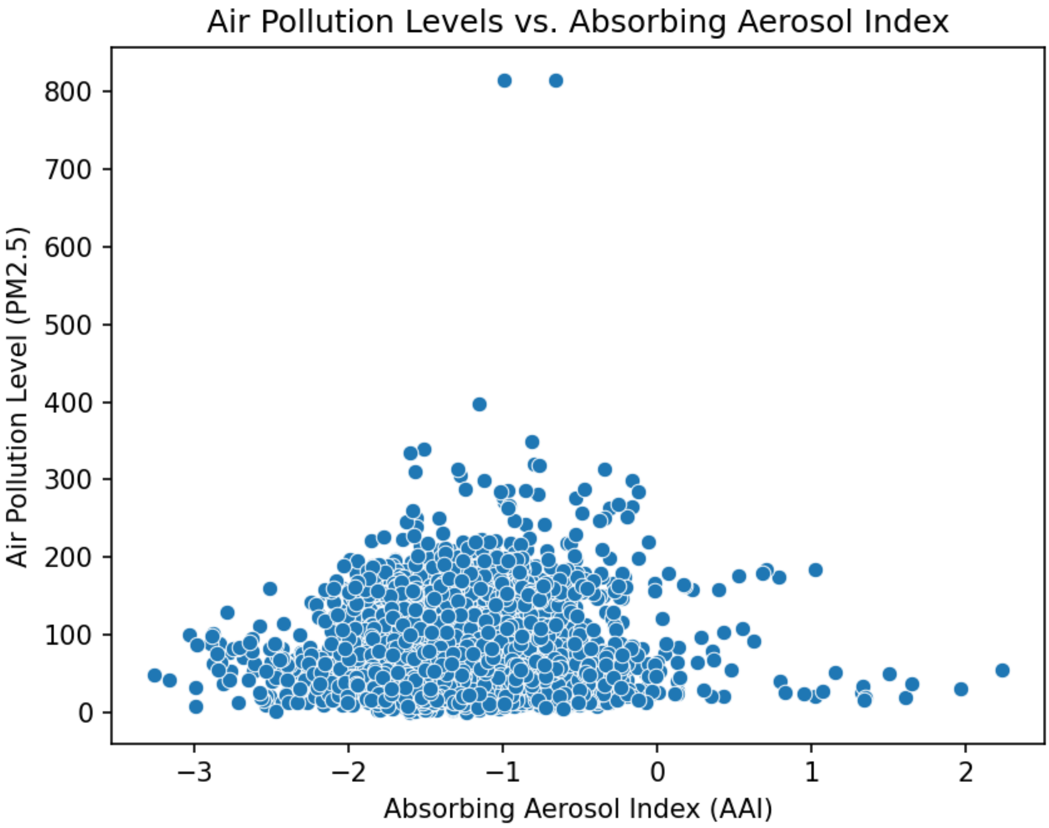

Figure 6: Air Pollution Levels vs. Absorbing Aerosol Index

The figure 7 depicts the relationship between the SO2 column number density and the air pollution level. The x-axis represents the SO2 column number density, and the y-axis represents the air pollution level.

The scatter plot displays a positive correlation between SO2 column number density and PM2.5 levels. Higher SO2 levels are associated with industrial activities and burning fossil fuels, which release sulfur dioxide into the air. Elevated SO2 concentrations contribute to air pollution and can result in increased PM2.5 concentrations. Based on the positive correlation, it can be concluded that higher SO2 column number density is likely to contribute to increased PM2.5 concentrations. Implementing measures to reduce SO2 emissions can help mitigate air pollution and potentially lower PM2.5 levels.

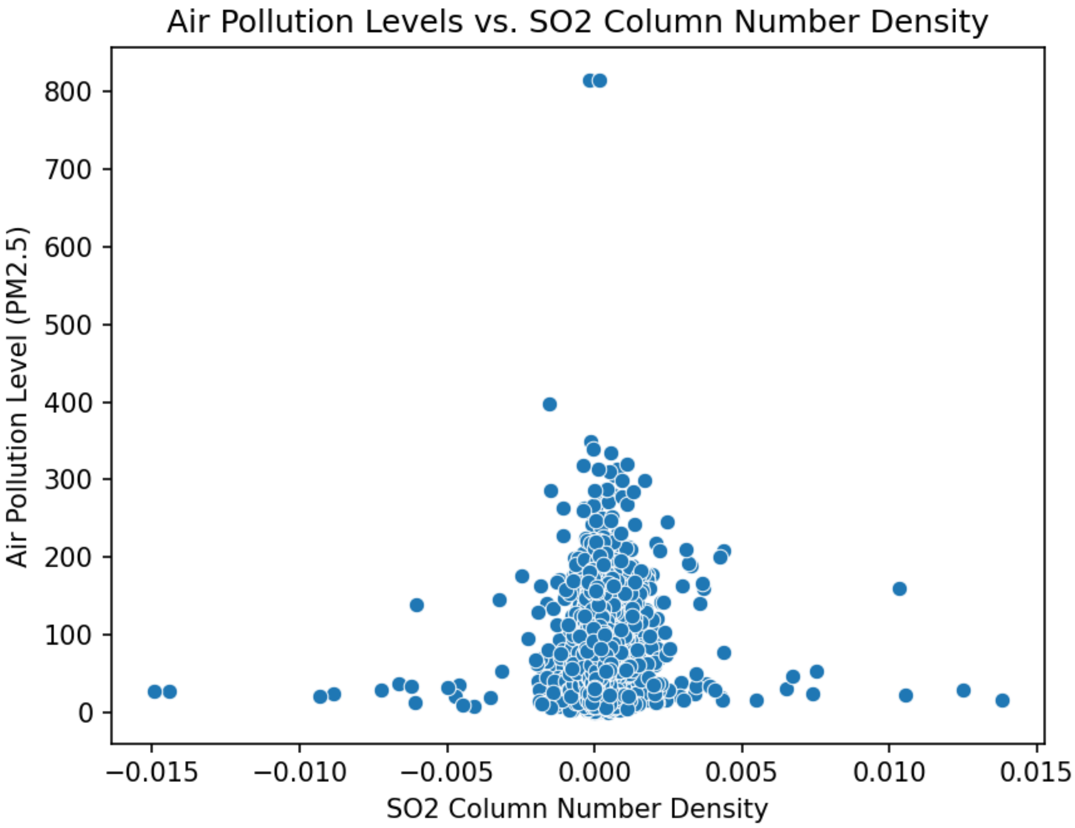

Figure 7: Air Pollution Levels vs. SO2 Column Number Density

6.3 Data Statistics

The Urban Air Pollution Challenge dataset, contains 30,557 rows and 82 columns. However, it's important to note that many of the sensor features in the dataset have null values. To address this issue, I performed data cleaning, after cleaning the data, I divided it into a training set, which comprised 80% of the data, and a testing set, which contained the remaining 20%. For the training data, I used a process known as "median imputation." This involves replacing the missing values with the median and for the testing data, I dropped all the missing values . It's mentioning that the testing data was , consisting of only 818 record after removing missing value while the training was 24445 record .

Additionally, it's crucial to note that certain features were dropped during the prediction process for both the training and testing data. These features included "Place_ID X Date," "Date," "Place_ID," and "target." The "Place_ID X Date" feature likely represented a combination of the unique identifier for a specific place (Place_ID) and the date of the air pollution measurement. The "Date" and "Place_ID" features served as identifiers and were not directly relevant to predicting air pollution levels. Finally, the "target" feature corresponded to the actual air pollution levels and was not included as a feature for prediction. Removing these features allowed the models to focus solely on the remaining sensor features, reducing noise and improving the accuracy of the predictions.

7 Section Seven: Research approach and methodologies

Quantitative Research Methodology: Quantitative research involves the collection and analysis of numerical data to understand patterns, trends, and relationships between variables. It aims to quantify phenomena and make statistical inferences about a population based on a sample. In quantitative research, data is typically collected through structured surveys, experiments, or other objective methods. Statistical analysis techniques are then applied to interpret the data and draw conclusions. The emphasis is on objectivity, generalizability, and statistical validity [14].

Qualitative Research Methodology: Qualitative research focuses on exploring and understanding complex phenomena in depth. It aims to uncover the meaning, experiences, and perspectives of individuals or groups. Qualitative research methods involve collecting non-numerical data, such as interviews, observations, and open-ended surveys. Researchers often employ techniques like thematic analysis, content analysis, or grounded theory to interpret the data and identify patterns, themes, or insights. The emphasis is on subjective interpretations, context, and richness of data [14].

For my research project, I have selected a quantitative research methodology 8 to collect and analyze air pollution data using IoT devices equipped with sensors. These sensors provide real-time and precise measurements of air quality parameters such as particulate matter (PM), gases, temperature, and humidity. The use of IoT devices allows for continuous data collection, enabling a comprehensive understanding of air pollution patterns.

The collected quantitative data is analyzed using Google Colab, a powerful data analysis and visualization tool. Google Colab facilitates the exploration and visualization of the data, allowing me to gain valuable insights into pollution patterns. The visualizations generated by Google Colab help in identifying correlations, trends, and outliers in the data, which can further inform decision-making and mitigation strategies.

In addition to data analysis, machine learning techniques are applied to the quantitative data to study and predict air pollution levels. By leveraging machine learning algorithms, patterns and trends within the data can be identified, enabling the development of predictive models. These models play a crucial role in forecasting future pollution levels and informing targeted interventions and mitigation strategies.

The selection of a quantitative research methodology is supported by its ability to provide objective and measurable data about air pollution. The use of IoT devices and sensors ensures accurate and continuous data collection, which is crucial for understanding the dynamic nature of air pollution. The application of Google Colab aids in data exploration and visualization, facilitating a comprehensive analysis of the collected data.

Overall, the combination of quantitative research methodology, IoT devices, data visualization using Google Colab, and machine learning techniques provides a robust framework for studying and predicting air pollution levels. This approach enables a comprehensive understanding of air quality and supports the development of targeted interventions and mitigation strategies to reduce air pollution.

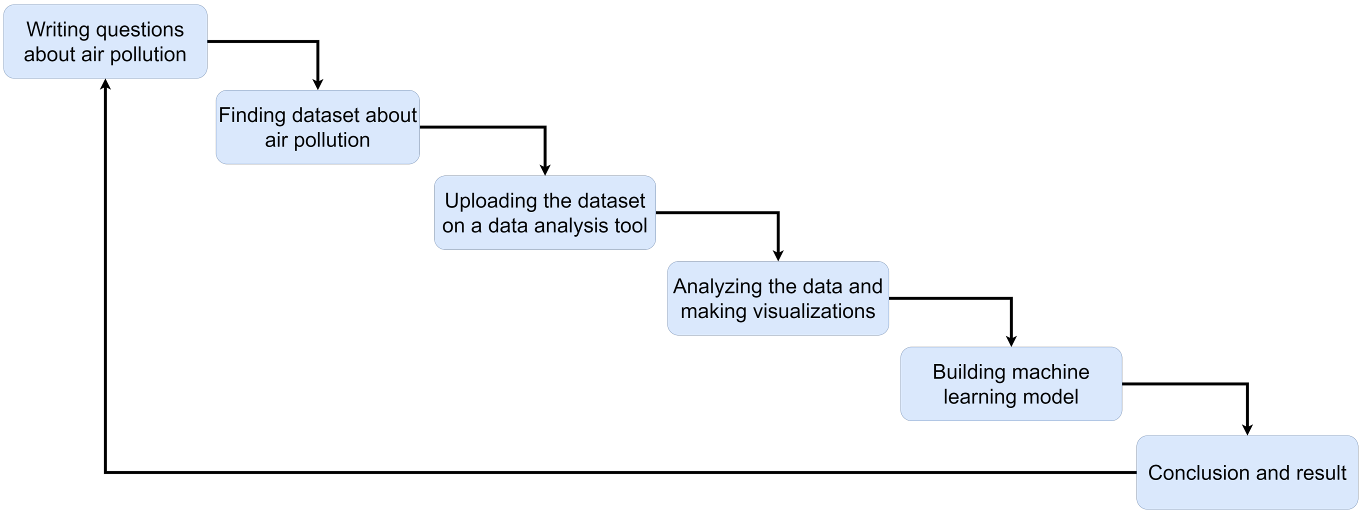
Writing questions about air pollution
Finding dataset about air pollution
Uploading the dataset on a data analysis tool
Analyzing the data and making visualizations
Building machine learning model
Conclusion and result

Figure 8: Project phases

How can we reduce air pollution?

The data collected through IoT devices provides valuable insights into the sources and patterns of air pollution. By analyzing this data, it becomes possible to identify the major contributors to air pollution and develop targeted strategies for reducing emissions from specific sectors or activities.

How can air pollution affect human life?

The data collected through IoT devices helps in monitoring and measuring air pollution levels in real-time. By studying this data, researchers can assess the correlation between air pollution levels and human health outcomes, such as respiratory illnesses, cardiovascular problems, and overall well-being. These findings contribute to a better understanding of the adverse effects of air pollution on human life.

How can IoT help to reduce and measure air pollution through data collection?

IoT devices equipped with sensors can continuously collect data on various air quality parameters, including particulate matter, gases, temperature, and humidity. This real-time data collection allows for more accurate and frequent monitoring of air pollution levels compared to traditional monitoring methods. By providing comprehensive and up-to-date information, IoT-enabled data collection supports proactive measures to reduce air pollution and enables timely interventions when pollution levels exceed safe thresholds.

How to use the data collected using IoT to train a machine learning model to reduce air pollution?

The data collected through IoT devices can be used to train machine learning models. By leveraging machine learning algorithms, patterns, and trends within the data can be identified, allowing for the development of predictive models. These models can forecast future pollution levels, identify potential pollution sources, and suggest effective mitigation strategies. By utilizing the IoT-collected data, machine learning models can aid in making informed decisions and implementing targeted interventions to reduce air pollution effectively.

7.1 Machine learning approaches and ensemble learning

1. **Machine learning approaches:**

   - Linear Regression: Linear Regression is a supervised machine learning algorithm that aims to establish a predictive relationship between independent variables and the outcome of an event. It seeks to find the best-fitting straight line that represents the data points, with the output being a continuous numerical value. For instance, it can be used to predict revenue, sales, or the number of products sold based on independent variables. The mathematical representation of linear regression is given by [15]: $y = 0 + 1x +$ In this equation, the dependent variable is denoted by y, the independent variable(s) by x, 0 represents the intercept of the line, 1 represents the linear regression coefficient (the slope of the line), and represents the random error. The presence of a random error is necessary as the line does not perfectly pass through all the data points.

     The linear regression model establishes a linear relationship between the dependent variable (y) and one or more independent variables (x). It quantifies how the value of the dependent variable changes in response to changes in the independent variable(s). The relationship is represented by a straight line with a slope, indicating the direction and magnitude of the change. The linear regression model is widely used for prediction and inference in various fields, providing valuable insights into the relationship between variables [15].

   - Random Forest Regression: A random forest is a powerful machine-learning technique that combines multiple classifiers to solve regression and classification problems effectively. It uses ensemble learning and consists of numerous decision trees trained through bagging. The algorithm aggregates tree predictions by averaging or taking the mean. Random forests overcome decision tree limitations, reducing over-fitting and providing reasonable predictions without extensive tuning. They offer improved accuracy, handle missing data effectively, and require minimal hyperparameter tuning. Decision trees are fundamental building blocks, forming a hierarchy with decision nodes and leaf nodes. Random forests randomly establish and segregate nodes, utilizing the bagging method. The final prediction is determined by a majority vote among the decision trees. Random forests offer robust and accurate solutions for complex problems, leveraging diverse perspectives from individual trees [16].

   - Gradient Boosting Regression: Gradient Boosting Regression is a powerful algorithm that solves regression problems in machine learning. It combines boosting and gradient descent to create an ensemble of weak learners that collaborate for accurate predictions. The algorithm begins with an initial prediction, often the target variable's average. It incrementally adds weak learners, like shallow decision trees, to the ensemble. Each new learner corrects the mistakes of previous learners. During each iteration, the algorithm computes the gradient of the loss function based on ensemble predictions. This gradient indicates the direction to adjust predictions and minimize the loss. The new learner is trained to predict the negative gradient, thus reducing the ensemble's error. Ensemble predictions are made by multiplying each learner's output by a learning rate, controlling their influence on the final prediction. A lower learning rate enhances robustness but slows down the algorithm. The ensemble prediction is obtained by summing all the learners' predictions. This process continues until a specified number of learners is reached or the loss function converges [17].

- Stacking Regression: Stacking Regression is an ensemble learning technique used for combining multiple regression models to make more accurate predictions. It leverages the concept of stacking, where the predictions of individual models are combined using another model called a meta-regressor. In Stacking Regression, a set of diverse base regression models is trained on the training data. These base models can be different regression algorithms or the same algorithm with varying hyperparameters. Each base model makes predictions on the validation or test set. The predictions of the base models are then used as input features for a meta-regressor, which is another regression model. The meta-regressor learns to combine the base models' predictions and produce the final prediction. It uses the base models' predictions as input to learning a higher-level relationship between the features and the target variable [18].

- Decision Tree Regression is a machine learning algorithm used for solving regression problems. It works by recursively partitioning the input feature space into distinct regions and assigning a constant value to each region. In decision tree regression, the algorithm builds a binary tree structure by repeatedly splitting the data based on the values of input features. Each internal node of the tree represents a splitting condition, and each leaf node represents a predicted value. The splitting process is based on minimizing the sum of squared errors (SSE) or another suitable error metric. At each step, the algorithm selects the best feature and split point that minimizes the error. This process is repeated recursively until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples per leaf [19].

2. **Ensemble learning:** Ensemble learning is a widely utilized approach in machine learning that combines multiple models to improve overall performance. It involves aggregating the predictions of individual models, known as base learners, to create a final prediction that often surpasses the capabilities of any single model. In the context of reliability analysis, existing ensemble-learning methods typically integrate ensemble learning with a learning function. One common strategy is to pre-construct the initial training set and test set. The training set is used to train the initial ensemble model, while the test set is employed to allocate weight factors and assess the convergence criterion [20].

   However, traditional approaches in reliability analysis primarily focus on local prediction accuracy near the limit state surface, rather than the global prediction accuracy across the entire space. Unfortunately, randomly generated samples in the initial training set and test set may result in the learning function's inability to identify the true "best" update samples. Additionally, the allocation of weight factors may be suboptimal or even unreasonable. These issues significantly impact the overall performance of the ensemble model [20].

   To address these challenges, the authors propose a novel hierarchical ensemble-learning framework (ELF) for reliability analysis. The ELF consists of two-layer models and encompasses three distinct phases. Within this framework, we introduce a new method called CESM-ELF, which integrates the classical ensemble of surrogate models (CESM) into the proposed ELF. Through extensive investigation across four examples, they demonstrate that CESM-ELF outperforms CESM in terms of prediction accuracy and exhibits improved efficiency in certain scenarios [20].

7.2 How the machine learning model can be used in IoT to reduce air pollution?

Machine learning models integrated into IoT systems offer the potential to predict air quality levels in real-time and provide alerts to individuals, organizations, and authorities, enabling them to take necessary actions to mitigate pollution. By combining machine learning algorithms with data collected from IoT sensors, these models can generate accurate and timely predictions of air quality, serving as valuable insights for proactive interventions [21].

To begin, strategically deployed IoT sensors continuously monitor various air pollutants such as particulate matter (PM), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), and other relevant indicators. These sensors collect data on pollutant levels, weather conditions, geographical factors, and contextual information, establishing a comprehensive dataset. The collected data is then utilized to train the machine learning model, leveraging historical air quality data. By analyzing this dataset, the model learns intricate relationships between different parameters and air quality levels. This training phase empowers the model to discern patterns, correlations, and trends, enabling it to make accurate predictions when provided with real-time sensor readings [21]. Once the model is trained, it becomes capable of generating predictions of air quality levels in real time. These predictions play a crucial role in identifying situations where air quality rapidly deteriorates or exceeds predefined thresholds, indicating the presence of poor air quality that necessitates immediate attention and intervention.

Organizations can greatly benefit from these predictions by integrating them into their systems or dedicated platforms, and receiving alerts that prompt them to take proactive measures. These alerts empower organizations to implement emission reduction strategies, optimize their operations, and adopt appropriate measures to minimize their impact on air quality. For instance, factories can adjust their production schedules, transition to cleaner energy sources, or employ emission control technologies in response to alerts regarding poor air quality [21]. Furthermore, authorities responsible for environmental management and public health can receive real-time alerts, enabling them to take swift and effective actions. These alerts facilitate the dissemination of public advisories, the implementation of traffic management strategies, and the allocation of resources toward addressing pollution sources. Leveraging these predictions, authorities can enhance their decision-making processes and allocate resources effectively to mitigate pollution, ultimately safeguarding public health [21].

7.3 Evaluation metrics

Evaluation metrics are used to measure the quality of a machine learning model. They are used to compare different models and to evaluate the performance of a model over time. there are many There are many different evaluation metrics available.In this project we used below Evaluation metrics.

- **R-squared: R-squared (R2):** is a statistical measure that quantifies the proportion of the variance in a dependent variable explained by an independent variable in a regression model. It goes beyond correlation by indicating how much the variance of one variable can explain the variance of another variable. For example, an R2 value of 0.50 means that approximately half of the observed variation can be explained by the model's inputs[22]. The equation for R-squared is as follows [22]:

  $$\mathbf{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  In this formula, represents the summation or summing up of values. The numerator represents the sum of squared errors (unexplained variance), and the denominator represents the total variance. Subtracting the numerator from 1 gives the proportion of the total variation that is explained by the independent variable, which is the R-squared value. When interpreting R-squared in the context of the dataset for urban air pollution in Zindi Africa, which was worked on with an 80 percent training and 20 percent testing split, a high R-squared value (85% to 100%) indicates a close alignment between the predicted air pollution levels and the actual observations in the dataset. On the other hand, a low R-squared value (70% or less) suggests that the predictions do not generally follow the variations in air pollution levels.

  A higher R-squared value can provide more confidence in the reliability of the beta figure, which is a measure of risk-adjusted returns. It's important to note that if the R-squared value is high but the beta is below 1, it may imply the potential for higher risk-adjusted returns in relation to urban air pollution predictions. It's worth mentioning that the dataset was divided into 80 percent for training and 20 percent for testing to assess the performance of the model and evaluate its predictive accuracy on unseen data. This split ensures a sufficient amount of data is used for training the model while reserving a portion for unbiased evaluation during testing.

- **Mean Squared Error (MSE):** The Mean Squared Error (MSE) is a statistical measure that assesses the proximity of a regression line to a set of data points. It represents the expected value of the squared error loss and is calculated by taking the average of the squared errors from the data concerning a function. A larger MSE indicates that the data points are widely dispersed around their mean, while a smaller MSE suggests that the data points are closely clustered around the mean. A smaller MSE is generally preferred as it signifies a more centralized distribution of data values, the absence of skewness, and fewer errors in terms of the dispersion of data points from the mean [23].

  In regression analysis, the MSE serves as an important indicator of the quality of the estimator. A smaller MSE corresponds to a smaller error, thus indicating a better estimator. It quantifies the accuracy of predictions made by the regression model. The calculation of MSE involves summing the squared differences between the actual data values and the corresponding predicted values, then dividing this sum by the sample size (n). The formula for MSE is as follows [23]:

  $$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

  In this formula, represents the summation or summing up of values, n denotes the sample size, "y" represents the actual data value, and "$\bar{y}$" represents the predicted data value. When performing regression analysis, MSE is particularly suitable when the target variable is assumed to follow a normal distribution and larger errors should be penalized more heavily than smaller ones. It provides a measure of the overall goodness of fit of the regression model [23].

- **Root Mean Squared Error (RMSE):** The Root Mean Squared Error (RMSE) is a key performance metric used in regression models to evaluate the accuracy of predictions. It represents the average difference between the predicted values and the actual values. The RMSE is an important indicator of how well the model can estimate the target variable [24]. A lower RMSE value indicates a better-performing model. In an ideal scenario, where the model predicts the exact expected values, the RMSE would be 0. One advantage of the RMSE is that it is expressed in the same unit as the predicted column, making it easily interpretable. For example, if the goal is to predict amounts in dollars, the RMSE can be directly understood as the amount of error in dollars. To reduce the RMSE, one can include additional influential variables in the training dataset. By incorporating more relevant features, the model can capture more nuances and improve its predictive performance [24].

The formula for calculating RMSE is as follows [24]:

$$\textbf{RMSE} = \sqrt{MSE}$$

7.4 Highest Performed Model

I have experimented with different machine learning algorithms to predict whether air pollution (PM2.5) will increase or not. These algorithms are Random Forest Regression, Stacking Regression, Linear Regression, Decision Tree Regression, and Gradient Boosting Regression. The best model in my experiments is Random Forest Regression. We have prepared the random forest regression model with n_estimators=100 and random_state=42 and we achieved the best score Mean Squared Error (MSE), R^2 Score, and RMSE Score, which are 735.09, 0.78, and 27.11, respectively.

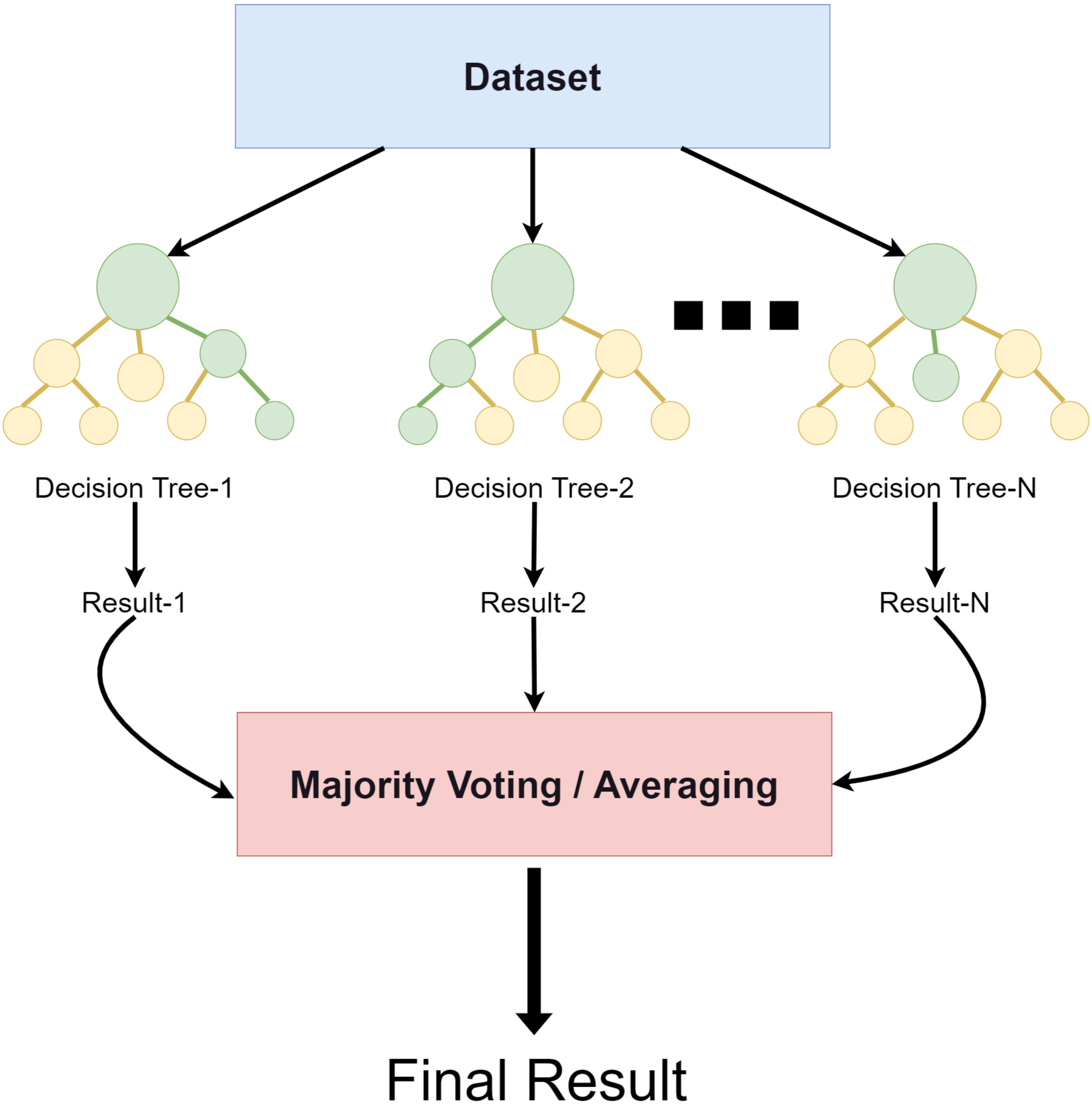

Figure 9: Proposed Model

8 Section Eight: Results and Discussion

For all ML algorithms, I have calculated the R-Squared, RMSE, MSE. This table 1 shows the evaluation results of the models. The models are in descending order from the highest accuracy to the lowest accuracy. The Random Forest Regression gained the highest accuracy with 78.92%, and Linear Regression got the most insufficient accuracy with 74.02%.

Based on the evaluation results of different machine learning algorithms 1, it can be observed that the Random Forest Regression model achieved the highest accuracy of 78.92% in predicting air pollution in Africa. This indicates a relatively strong relationship between the independent variables and the dependent variable. Furthermore, the R-Squared values for all the models range from 0.7402 to 0.7892, suggesting that the selected independent variables explain approximately 74.02% to 78.92% of the variance in air pollution levels. These models also exhibit relatively low Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) values, indicating a good fit between the predicted and actual air pollution values. Considering these results, it is reasonable to conclude that the developed machine learning models have the potential to predict future air pollution levels in Africa with a certain degree of accuracy. However, further analysis and continuous monitoring are necessary to assess the specific factors influencing air pollution and to make more precise predictions regarding the future increase or decrease of air pollution in the region.

Table 1: RESULTS OF DIFFERENT MACHINE LEARNING ALGORITHM

Model	**R-Squared**	**RMSE**	**MSE**
Random Forest Regression	0.7892	27.11	735.09
Gradient Boosting Regression	0.7843	27.42	752.15
Stacking Regression	0.7743	28.21	762.11
Decision Tree Regression	0.7463	29.74	884.60
Linear Regression	0.7402	30.09	905.69

9 Section Nine: Conclusion

In conclusion, this research project addresses the pressing issue of air pollution and proposes a digital transformation approach using IoT and machine learning. The problem of air pollution (PM2.5), particularly in Africa, is a significant challenge that affects human health, ecosystems, and overall well-being. The study aims to utilize IoT data collection and analysis, along with machine learning techniques, to find effective solutions for reducing air pollution. Through the collection of data from IoT devices, valuable insights into the sources and patterns of air pollution have been obtained. By analyzing this data, it becomes possible to identify major contributors to air pollution and develop targeted strategies for emission reduction. Furthermore, the correlation between air pollution levels and human health outcomes has been explored, highlighting the detrimental effects on respiratory illnesses, cardiovascular problems, and overall quality of life.

IoT devices equipped with sensors play a crucial role in measuring and monitoring air pollution in real time. This continuous data collection enables more accurate and frequent assessments of pollution levels compared to traditional methods. By providing comprehensive and up-to-date information, IoT-enabled data collection supports proactive measures to reduce air pollution and facilitates timely interventions when pollution levels exceed safe thresholds. Machine learning models trained on the IoT-collected data offer valuable predictive capabilities for addressing air pollution. Through the application of machine learning algorithms, patterns and trends within the data can be identified, leading to the development of predictive models. These models can forecast future pollution levels, identify potential pollution sources, and suggest effective mitigation strategies. Leveraging IoT data, machine learning models enable informed decision-making and targeted interventions to effectively reduce air pollution.

In this research project, the Random Forest Regression model emerged as the most successful in predicting air pollution levels. With its high accuracy rate, the model demonstrates a strong relationship between independent variables and air pollution. The R-Squared values, along with low RMSE and MSE scores, validate the model's ability to explain a significant portion of the variance in air pollution levels and make accurate predictions.

Overall, this research highlights the potential of digital transformation through IoT and machine learning approaches in addressing air pollution. By leveraging IoT data collection, data analysis using tools like Google Colab, and machine learning techniques, it becomes possible to gain comprehensive insights into air pollution patterns and develop targeted interventions. However, further analysis, continuous monitoring, and exploration of specific factors influencing air pollution are necessary to refine predictions and implement effective long-term solutions. The findings of this study provide a foundation for future work in utilizing IoT and machine learning in environmental management to combat the pervasive issue of air pollution.

References

[1] S. Mosley, "Environmental history of air pollution and protection," *The basic environmental history*, pp. 143–169, 2014.

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

[3] A. K. Feroz, H. Zo, and A. Chiravuri, "Digital transformation and environmental sustainability: A review and research agenda," *Sustainability*, vol. 13, no. 3, p. 1530, 2021.

[4] ——, "Digital transformation and environmental sustainability: A review and research agenda," *Sustainability*, vol. 13, no. 3, p. 1530, 2021.

[5] S. Chowdhury, M. S. Islam, M. K. Raihan, and M. S. Arefin, "Design and implementation of an iot based air pollution detection and monitoring system," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*. IEEE, 2019, pp. 296–300.

[6] ——, "Design and implementation of an iot based air pollution detection and monitoring system," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, 2019, pp. 296–300.

[7] N. S. Desai and J. S. R. Alex, "Iot based air pollution monitoring and predictor system on beagle bone black," in *2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, 2017, pp. 367–370.

[8] S. Dhingra, R. B. Madda, A. H. Gandomi, R. Patan, and M. Daneshmand, "Internet of things mobile–air pollution monitoring system (iot-mobair)," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5577–5584, 2019.

[9] D. Munera, J. Aguirre, N. G. Gomez *et al.*, "Iot-based air quality monitoring systems for smart cities: A systematic mapping study," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, p. 3470, 2021.

[10] O. Alruwaili, I. Kostanic, A. Al-Sabbagh, and H. Almohamedh, "Iot based: Air quality index and traffic volume correlation," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2020, pp. 0143–0147.

[11] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC public health*, vol. 17, pp. 1–19, 2017.

[12] A. Zbiciak and T. Markiewicz, "A new extraordinary means of appeal in the polish

criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment," *Access to Justice in Eastern Europe*, vol. 6, no. 2, pp. 1–18, Mar. 2023.

[13] Zindi. [Online]. Available: https://zindi.africa/competitions/zindiweekendz-learning-urban-air-pollution-challenge/data

[14] P. Aspers and U. Corte, "What is qualitative in qualitative research," *Qualitative sociology*, vol. 42, pp. 139–160, 2019.

[15] "Linear Regression in Machine Learning [with Examples] — knowledgehut.com," https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning, [Accessed 10-Jun-2023].

[16] "Introduction to Random Forest in Machine Learning — section.io," https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/, [Accessed 10-Jun-2023].

[17] A. Saini, "Gradient Boosting Algorithm: A Complete Guide for Beginners — analyticsvidhya.com," https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/, [Accessed 10-Jun-2023].

[18] Y. Khandelwal, "Ensemble Stacking for Machine Learning and Deep Learning — analyticsvidhya.com," https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/, [Accessed 10-Jun-2023].

[19] "How Decision tree classification and regression algorithm worksx2014;ArcGIS Pro | Documentation — pro.arcgis.com," https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-decision-tree-classification-and-regression-works.htm, [Accessed 10-Jun-2023].

[20] C. Zhou, H. Zhang, M. A. Valdebenito, and H. Zhao, "A general hierarchical ensemble-learning framework for structural reliability analysis," *Reliability Engineering & System Safety*, vol. 225, p. 108605, 2022.

[21] T. W. Ayele and R. Mehta, "Air pollution monitoring and prediction using iot," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1741–1745.

[22] "R-Squared: Definition, Calculation Formula, Uses, and Limitations — investopedia.com," https://www.investopedia.com/terms/r/r-squared.asp, [Accessed 10-Jun-2023].

[23] "Mean Squared Error : Overview, Examples, Concepts and More | Simplilearn — simplilearn.com," https://www.simplilearn.com/tutorials/statistics-tutorial/mean-squared-error, [Accessed 10-Jun-2023].

[24] "SAP Help Portal — help.sap.com," https://help.sap.com/docs/SAP_PREDICTIVE_ANALYTICS/41d1a6d4e7574e32b815f1cc87c00f42/5e5198fd4afe4ae5b48fefe0d3161810.html, [Accessed 10-Jun-2023].