

GLOBAL AIR POLLUTION PREDICTION

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

ASHISH KUMAR (Reg.No - 40110125)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Category-1 University by UGC

Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

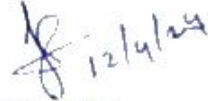
CHENNAI - 600119

April - 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

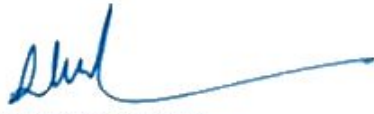
BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **ASHISH KUMAR** (Reg. No. 40110125) who carried out the Project entitled "**GLOBAL AIR POLLUTION PREDICTION**" under my supervision from November 2023 to April 2024.



Internal Guide

Dr. J. JESLIN SHANTHAMALAR, M.E., Ph. D.



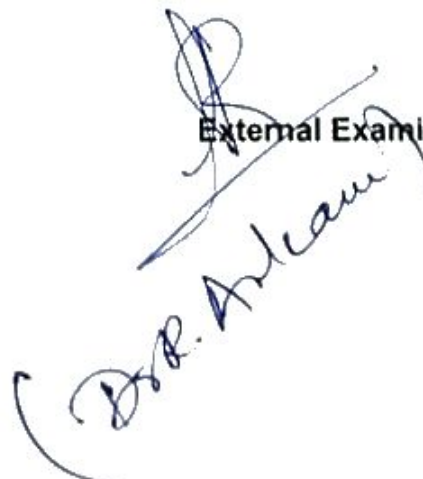
Head of the Department
HEAD OF THE DEPARTMENT
Dr. DEEPAKSHIMANAN, M.E., Ph.D.
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Jeppiaar Nagar, Rajiv Gandhi Salai,
Chennai - 600 119.

Submitted for Viva voce Examination held on 15-04-24


Internal Examiner

(V SARANYA)

External Examiner


(Dr. Arulam)

DECLARATION

I, **ASHISH KUMAR** (Reg.No - 40110125), hereby declare that the Project Report entitled "**GLOBAL AIR POLLUTION PREDICTION**" done by me under the guidance of **Dr. J. JESLIN SHANTHAMALAR, M.E., Ph.D.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 12-04-2024



PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **The Board of Management of Sathyabama Institute of Science and Technology** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph. D, Dean**, School of Computing, and **Dr. L. Lakshmanan M.E., Ph. D.**, Head of the Department, Dept. of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. J. Jeslin Shanthamalar, M.E., Ph. D.**, for her valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Air pollution is a worldwide environmental concern that significantly impacts the health and well-being of individuals across the globe. The impact of air pollutants on the health of human beings, ecosystems, and the climate has spurred significant concern and necessitated comprehensive analysis and understanding. This study provides an in-depth examination of global air pollution trends, sources, and their consequences, drawing from a vast and diverse dataset. The "Global Air Pollution Dataset" serves as the foundation for this analysis, offering a wealth of information that encompasses air quality measurements, emissions data, meteorological variables, and more. The dataset combines historical records with real-time updates, providing a rich resource for researchers, policymakers, and environmental enthusiasts. The proposed model achieved 98% in terms of accuracy using Decision Tree Algorithm to analyze the Pollutants as a public health concern that include particulate matter, carbon monoxide, nitrogen dioxide and ozone. Air pollution is a big problem that affects everyone around the world. It happens when harmful stuff gets into the air we breathe from things like factories, cars, farming, and natural events like wildfires. This stuff includes tiny particles, gases, and chemicals that can make us sick and harm the environment. Air pollution makes our air dirty and can cause health problems like breathing issues and heart disease. It also damages nature by causing acid rain and harming plants and animals. Plus, it contributes to climate change, making the Earth hotter. To fix air pollution, countries need to work together. They can make rules to limit pollution from factories and cars and invest in cleaner energy like solar and wind power. Scientists are also working on new technologies to help reduce pollution. It's important for everyone to take action to clean up the air we breathe. By working together, we can make our world healthier and safer for everyone.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	i
	LIST OF TABLES	iv
	LIST OF FIGURES	v
	LIST OF ABBREVIATIONS	vi
1	INTRODUCTION	1
	1.1 Motivation	2
	1.2 Objective of the Project	3
	1.3 Project Explanation	5
2	LITERATURE SURVEY	6
	2.1 Inferences from Literature Survey	12
	2.2 Open problems in Existing System	13
3	REQUIREMENTS ANALYSIS	15
	3.1 Aim of the Investigation	15
	3.2 Scope of the Investigation	15
	3.3 Feasibility Studies of the Project	16
	3.3.1 Project Description and Objectives	16
	3.3.2 Technical Feasibility	17
	3.3.3 Financial Feasibility	17
	3.3.4 Operational Feasibility	18
	3.3.5 Organizational Feasibility	18
	3.3.6 Environmental and Social Impact	19
	3.4 Risk Analysis of the Project	19
	3.5 Software Requirements Specification Document	20
4	DESCRIPTION OF PROPOSED SYSTEM	21
	4.1 Selected Methodology or process model	21
	4.1.1 Data Mining and Machine Learning	22

	4.1.2	Air Quality Modeling	22
	4.1.3	Risk Assessment and Health Impact Analysis	23
	4.2	Architecture / Overall Design of Proposed System	25
	4.2.1	Data Collection	25
	4.2.2	Data Processing	26
	4.2.3	Data Augmentation	26
	4.2.4	Feature Selection	26
	4.2.5	Model Training	27
	4.3	Description of Software for Implementation and Testing plan of the Proposed Model/System	28
5		IMPLEMENTATION DETAILS	29
	5.1	Development and Deployment Setup	29
	5.2	Algorithms	31
	5.2.1	Testing	32
6		RESULTS AND DISCUSSIONS	34
	6.1	Results	34
	6.2	Discussion	35
	6.2.1	Points to be Checked and Clear	35
	6.3	Performance Analysis	36
7		CONCLUSION	38
	7.1	Summary	38
	7.2	Conclusion	38
		REFERENCES	40
		APPENDIX	43
	A.	SOURCE CODE	43
	B.	SCREENSHOT	52
	C.	RESEARCH PAPER	56

LIST OF TABLES

TABLE NO.	TABLE NAME	PAGE NO.
2.1	Survey on existing System	13
5.1	Data based on the Country	30
5.2	Comparision between two Algorithms	32
5.3	Training Data Result	32
5.4	Testing Data Result	33
6.1	Testing Data for Accuracy	34
6.2	Comparision between the Algorithms	35

LIST OF FIGURES

FIG. NO.	FIGURE NAME	PAGE NO.
4.1	Air Quality for Maximum values of AQI	23
4.2	Risk of Health Impact on Varius Places	24
4.3	System Architecture for Global Air Pollution Prediction	25
5.1	Density of Polluted Air	30
5.2	Well Represented Country	31

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
AQI	Air Quality Index
BP	Back Propagation
CO	Carbon Monoxide
IDE	Integrated Development Environment
KNN	K-nearest neighbor
PM	Particulate Matter
RAM	Random Access Memory
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support vector Machine
SVR	Systemic vascular resistance

CHAPTER 1

INTRODUCTION

Air pollution is a pressing global environmental challenge with far-reaching consequences for public health, ecosystems, and climate change. The presence of pollutants such as particulate matter, nitrogen oxides, sulfur dioxide, carbon monoxide, and volatile organic compounds in the atmosphere has led to a host of adverse effects, including respiratory diseases, reduced air quality, and altered weather patterns.

Our eyes can see the milky way galaxy without any equipment's it didn't ends here animals and insects like beetles used to travel in the night sky straight without any hesitation. They were following a galaxy white strip-like structure. These days we can't be able to see because of pollution and it causes some myopia disease which affects human's eyes as well as animals.

The air pollution is multifaceted and complex, encompassing a spectrum of health effects, environmental degradation, and socioeconomic implications. Exposure to polluted air is linked or we can say that lead to health problems, ranging from respiratory diseases and lungs cancer to neurological damages.

Global air pollution prediction involves a multifaceted process with several key components. First and foremost, it relies on the collection of extensive data from various sources, including air quality monitoring stations, satellite measurements, and emissions data. These diverse datasets require thorough preprocessing, including data cleaning and spatial alignment, to ensure data quality and consistency.

Geospatial data representation is essential, as it enables the translation of geographical information into grid-like formats suitable for analysis. Machine learning models, often using techniques such as Decision Tree and Random Forest, play a crucial role in predicting air pollution levels, taking into account features like meteorological variables, geographical characteristics, emissions data, and

historical air quality values.

Model training and hyper parameter tuning are critical steps, and evaluation is performed using relevant metrics. Post-processing techniques can further enhance predictions, and visualizations are created to communicate results effectively. Ultimately, the deployment of the model for real-time or batch predictions and continuous improvement through data updates and model adaptation are essential to providing accurate and actionable information for global air pollution forecasting.

To enhance the real-world application of machine learning algorithms. Global air Pollution prediction scenario is taken. In industrialization, urbanization and transportation has led to an increase in the emission of pollutants in the atmosphere and making it a global critical issue.

Global air pollution prediction involves the development of advanced models and data-driven systems that can forecast air quality conditions worldwide. predictive models utilize a wealth of data, including historical air quality measurements, meteorological information, geographical factors, and other variables to provide real-time and future estimates of air pollution levels.

The increasing prevalence of air pollution has emerged as a pressing global concern, significantly impacting public health, ecosystems, and climate patterns. Anthropogenic activities, such as industrialization, transportation, and agriculture, have led to the release of pollutants into the atmosphere, contributing to the degradation of air quality. As societies become more interconnected and reliant on various sources of energy and production, the need for accurate and timely prediction of global air pollution has become imperative.

1.1 MOTIVATION

The motivation to tackle global air pollution arises from various concerns, including public health, environmental conservation, economic impacts, social justice, international cooperation, technological innovation, and policy regulation. Recognizing the adverse effects of air pollution on human health, ecosystems, and

economies, there's a collective drive to implement measures aimed at reducing pollution levels. This involves leveraging technological advancements, fostering international collaboration, and enacting effective policies and regulations. By addressing air pollution comprehensively and collaboratively, we aim to create a healthier, more sustainable future for all.

The motivation to address global air pollution is driven by various concerns, including public health, environmental conservation, economic impacts, social justice, international cooperation, technological innovation, and policy regulation. Recognizing the detrimental effects of air pollution on human health, ecosystems, economies, and social equity, there is a collective effort to implement measures aimed at reducing pollution levels. This involves leveraging technological advancements, fostering international collaboration, and enacting effective policies and regulations. By addressing air pollution comprehensively and collaboratively, we aim to create a healthier, more sustainable future for all, where clean air is accessible to everyone.

1.2 OBJECTIVE OF THE PROJECT

The project on global air pollution aims to comprehensively address the issue of air pollution on a worldwide scale. Its objectives include identifying sources of pollution, assessing its impacts, developing mitigation strategies, enhancing monitoring systems, promoting policy reform, engaging stakeholders, fostering international collaboration, and ensuring equity and justice. By pursuing these objectives, the project aims to contribute to protecting public health, preserving ecosystems, mitigating climate change, and promoting sustainable development globally.

The project seeks to identify and analyze the various sources of air pollution, including industrial emissions, vehicular exhaust, agricultural activities, biomass burning, and natural events like wildfires and volcanic eruptions. Understanding the sources allows for targeted interventions to mitigate pollution effectively.

It involves evaluating the multifaceted impacts of air pollution on human health, ecosystems, climate, and the economy. This includes quantifying the health risks associated with exposure to pollutants, assessing ecological damage, estimating

economic costs, and understanding the contribution of air pollution to climate change.

The project aims to develop and implement strategies to reduce air pollution levels. This may include promoting the adoption of clean energy technologies such as renewable energy sources and energy-efficient practices, improving transportation infrastructure to reduce emissions, implementing pollution control measures in industries, and promoting sustainable land-use practices.

Strengthening air quality monitoring networks and systems is essential for accurately tracking pollution levels, identifying pollution hotspots, and providing timely data for decision-making. This involves deploying advanced monitoring technologies, expanding monitoring coverage to underserved areas, and improving data sharing and accessibility.

The project aims to engage with stakeholders, communities, and the public to raise awareness about the impacts of air pollution and empower individuals to take action. This may involve organizing educational campaigns, community workshops, and outreach programs to promote behavioral changes, such as reducing personal exposure to pollutants and supporting clean air initiatives.

Facilitating collaboration and knowledge-sharing among countries, organizations, and researchers is essential for addressing transboundary pollution issues and implementing best practices for air quality management. This may involve establishing partnerships, sharing data and expertise, and supporting joint research initiatives to tackle common air pollution challenges.

The project aims to ensure that pollution reduction efforts prioritize the needs of vulnerable populations and address environmental injustices. This includes advocating for policies and interventions that promote equitable access to clean air and environmental resources, address disparities in exposure to pollution, and promote social justice in environmental decision-making.

1.3 PROJECT EXPLANATION

The project on global air pollution is aimed at addressing the pervasive effects of air pollution worldwide. With an approach on various aspects of pollution management, public health, environmental conservation, policy advocacy, and international collaboration, the project seeks to mitigate pollution levels and promote a cleaner, healthier, and more sustainable future for all.

At the core of the project is the identification of pollution sources. Through extensive research and analysis, the project endeavors to pinpoint the sources of air pollution, ranging from industrial emissions and transportation exhaust to agricultural activities and natural phenomena such as wildfires and volcanic eruptions. Understanding these sources is essential for developing targeted strategies to reduce pollution effectively.

Once the sources are identified, the project conducts comprehensive assessments to understand the full extent of air pollution's impacts. This involves quantifying health risks, evaluating ecological damage, estimating economic costs, and assessing the contribution to climate change. By assessing the impacts of pollution, the project can prioritize interventions and allocate resources effectively.

Strengthening air quality monitoring networks and systems is another critical aspect of the project. By deploying advanced monitoring technologies, expanding coverage to underserved areas, and improving data accessibility, the project aims to accurately track pollution levels and provide timely data for decision-making.

Collaboration among countries, organizations, and researchers is critical for addressing transboundary pollution issues. The project facilitates partnerships, data sharing, and joint research initiatives to tackle common air pollution challenges on a global scale.

CHAPTER 2

LITERATURE SURVEY

Doreswamy et al. (2020) investigated the ML predictive models for forecasting Particulate Matter concentration in the air. He studied six years of air quality and monitored data in Taiwan and applied for the models. According to them, pollution is increasing because of industrialization, urbanization, automobiles, power plants or chemicals, etc. They claimed that predicted values and actual values were very close to each other.

Madan T, Sagar S, Virmani D (2020) Air pollution is dangerous for human health and should decrease fast in urban and rural areas so it is necessary to predict the quality of air accurately. There are many types of pollution like water pollution, air pollution, soil pollution etc. The air pollution can seriously cause diseases like lung cancer, brain disease and even it can lead to death.

Madhuri VM, Samyama GGH, Kamalapurkar S (2020) The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. These can also be able to predict by using probability and statistics but these are the methods which are too complex to use so in place of these methods machine learning is better to use. The proposed work of using different algorithms such as LR, SVM, DT and RF. According to him the results show that the AQI predictions obtained through RF are promising.

C. Wu, W. Hu, M. Zhou, S. Li, and Y. Jia, (2019) To control and improve air quality, China's Ministry of Ecological Environment issued the "Ambient air quality standards" (GB 3095-2012) in February 2012, which set SO₂, NO₂, CO, O₃, PM₁₀ and PM_{2.5} as the basic monitoring indicators. He was trying to show that air pollution has gone through several stages in the evolution of single pollutants into composite pollutants which had been caused by the existence of many or several pollutants. These indicators are multidimensional, and they have profound impacts on evaluating air quality.

Bhalgat P, Bhoite S, Pitare S (2019) Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. According to him, global warming, acid rains, or an increased number of patients. With this increasing air pollution, He is implementing models which will record information about concentrations of air pollutants (SO₂, NO₂, etc). The deposition of this harmful gasses in the air is affecting the quality of people's lives, especially in urban areas.

G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie (2018) Monitoring and preserving air quality has become one of the most essential activities in many industrial and urban areas today. The quality of air has been affected due to various forms of pollution which has been caused by the transportation, electricity, fuels use etc. The deposition of harmful gasses is creating a serious threat like short term or long-term health effects in the cities. With increasing air pollution, they worked and implemented air quality monitoring models which collect information about the air pollutants and provide results of the air pollution in each area.

J. Wang, X. Zhang, Z. Guo, and H. Lu (2017) Air quality has received continuous attention from both environmental managers and citizens. Their early-warning systems for air pollution are very useful for avoiding negative health effects. Developing an early-warning system is very challenging, and also necessary. They decided to develop an early-warning system that will consist of a hybrid forecasting method for air quality prediction and assessment modules.

Y. Chen, L. Wang, F. Li, B. Du, K.-K. R. Choo, H. Hassan, and W. Qin (2016) Air quality has a significant influence on people's health. Air pollution can cause many respiratory diseases, whether good or best air quality is beneficial for physical and mental health. While inhaling polluted or dirty air for a severe long time that can cause many physical and psychological problems like difficulty in concentration or asthma. that can make people suffer while in traffic or travelling, which makes it use a large number of resources in the training process.

K. Kujaroentavon, S. Kiattisin, A. Leelasantitham, and S. Thammaboosadee (2014) air quality by the algorithm of decision tree using air quality index(AQI) in Thailand for pollutants is named as O₃, NO₂, CO, SO₂, PM₁₀ and levels of healthy concern. The purpose of this research is to establish rules on the basis of air quality classification by levels of health concerns. The results of this study were correctly classified into instances of a training set of 96.80% and the testing set of 91.07%. The algorithm of the decision tree can be used to become rules of the separated air quality classification by levels of health concern.

W. Bin (2008) air pollution in China, we studied the spatial-temporal characteristics of air pollution in China from 2001–2014. First, we drew several line charts and histograms for the high level clustering trends in some north cities and for the low level of the clustering southern cities of the Air Pollution and the Air Quality Index (AQI) of 31 capitals, their cities and municipalities to find the distribution across different times and cities.

Ramani RG, Shanthamalar JJ (2022) Jul. Automated image quality appraisal through partial least squares discriminant analysis paper presents a new feature extraction of methods which has been using full-reference and no-reference image quality metrics for the image quality classification using different machine learning algorithms.

K. Veljanovska and A. Dimoski(2018) found that the pollution Affecting human's respiratory systems is the cause for increased mortality and increased risk for diseases for the population. Many efforts from both local and state governments had been done in order to understand and predict the air quality index aiming to improve public health. This paper is one of the scientific contributions towards this challenge.

G. Parthasarathy and B. N. Chatterji (1990) Here they used the K-nearest-neighbor (KNN) estimates that give unbiased estimates of the probability density function of a random variable from observations of that of the random variables when K, the number of nearest neighbors. The total number of observations available tends to lead to infinity such that $K/N \rightarrow 0$. It is shown that in small sample problems they

give closer estimates to the true probability density than the traditional KNN estimates.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann (2010) as it is based on the precision recall curve which is totally based on classification performance. It is basically used in the performance on information retrieval or checks the overall performance. On the basis of accuracy on two limitations. Its first is if they have tested an imbalance dataset then its accuracy must have to be greater than 50 percent. The second is it will not see the difference between the error that they had got.

F. A. Gers (2001) as they have been trained in their neural network to predict a clean speech if there is some noise available in their background so they used a noise feature. Their System has two methods of DE reverberation: it means to clear the noise which is detected by microphones. The first method is Phase error-based filtering which performs on the basis of time difference means that the arrival of speech may be up and down. The second method is Correlation shaping. This speech is recorded by the help of a multi-microphone tablet device.

M. Zhao and X. Li (2011) as this is based on the air quality but they have distributed in a way of smart city. It means categorizing on the basis of people's livelihood, their environmental protection, public safety, etc. By making a weather smart grid it will improve overall understanding of the environment. It will help to get ready for any kind of climatic or weather change. These are the things which will help to keep the weather forecast accurately. Here they have used decision trees, decentralization or weighting.

X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang (2015) as the air mass trajectory was used here to recognize a separate corridor for transport of the dirty air means polluted air and transport of the clean air it means fresh air to a selected place. For this wind speed and direction also have been considered as parameters in calculating the trajectory of the polluted air to see the indication of that air. They found that the trajectory based on the geographical model is suitable for this method.

So their approach will show that it can be applied to other countries as well.

V. M. Niharika and P. S. Rao (2014) as they ran a survey on the air quality forecasting techniques they used the substance which is present in the atmosphere in a way that can harm humans, plants or animals and the environment will be damaged by those substances and it will show the effect on the weather or on the sudden climatic change. For analyzing these kinds of data they decided to use the Artificial neural network, Logics, and some software related to the ANN. For reading those harmful things they decided to take only some different pollutants.

D.Mishra (2016) as they used the Neuro-Fuzzy model at that time so for the air quality forecasting they chose to deal with the Nitrogen dioxide (NO₂) this is the pollutant that had been chosen for the analysis. That time the variables were temperature, pressure, humidity, wind speed and direction etc. They decided to first analyze the urbanized areas to start their analysis because of the growing places. As for analysis they also decided to use artificial neural networks and the Statistical measure.

T. Lan, Y. Zhang, C. Jiang, G. Yang, and Z. Zhao (2018) as they decided to work on the disease which is called as spread F they used to store the data in the format of ionograms so they decided to train and process to learn about this by using the decision tree. Ionogram is a set or we can say it as a collection. They took an ionogram which is an image and projected it to use the decision tree. By using an ionogram with spread F they got the accuracy which reached up to 89%, which is a good approach.

Xiaosong Zhao, Rui Zhang, Jheng-Long Wu (2018) as they used deep learning to go for the prediction of the air quality classification, they decided to use the Support vector machine, Random forest and the RNN model they used three different industrial places of United states on the basis of some specific pollutants like PM_{2.5}, PM₁₀, ozone etc. Among these algorithms they decided to choose RNN because it is giving the best accuracy for the air quality classification.

Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth and Hari Kiran Reddy (2020) as it used to analyze the air quality so they used air quality Index for only India to and used a useful approach for resolving the quality analysis of air. So on that basis they used many algorithms to analyze the best way to get better accuracy so they used Linear regression, Random Forest, Back Propagation, decision tree, Root mean square so that he can compare the data of its value to see which way is better.

Soubhik Mahanta, T. Ramakrishnudu, Rajat Raj Jha and Niraj Tailor (2019) as they decided to do this prediction for because they don't want any of the people to face danger or suffer from that so they decided to analyze urban areas or cities like Gurugram, Faisalabad, Delhi, etc. They choose those cities because they are the most polluted cities. So, if they do this prediction with their AQI values they can show them what kind of difficulties you are going to face in future. So they perform regression analysis on the data.

Zhili Zhao, Jian Qin, Zhaoshuang He, Huan Li, Yi Yang and Ruisheng Zhang (2020) as they decided to do the air quality prediction in the northwest of china. They decided to check for every hour, which means they will show the hourly prediction which means they have a huge impact on weather or pollution level and it can increase at any time so for that purpose they had decided to do hourly prediction. They used some of the substances or gasses to take measure on their prediction that pollutants are CO, NO₂, PM_{2.5}, etc. They used Serial data and neural networks for their predictive values.

Kostandina Veljanovska and Angel Dimovski (2018) as I understand in his paperwork, he is using AI tools so that it can make decisions on its own not by a programmer that will feed all the data and say it to show the values. He wants to say that after training the model it will have to predict on its own. So here he used some of the machine learning algorithms like neural network.

2.1 INFERENCES FROM LITERATURE SURVEY

In the field of air quality evaluation, there are many prediction methods. The most direct and efficient method is to calculate the air pollution index (API) according to the concentration value of designated air pollutants. Air quality is directly obtained through the analysis of the pollution index. This method is suitable for the analysis of air quality status and changing trends in the short term.

Air quality evaluation has been conducted using conventional approaches for many years. The traditional approaches for air quality prediction use mathematical and statistical techniques. In these techniques, a physical model is initially designed, and data is coded with mathematical equations. However, these are complex mathematical calculations. In addition, these methods provide limited accuracy. More recently, alternatives to traditional methods have been proposed; these alternatives use big-data and machine learning approaches. Many researchers have developed or used big-data-analytical models and machine-learning based models to conduct air quality evaluation to achieve better accuracy in evaluation and prediction.

There are several common algorithms used for air quality evaluation, such as the air quality index method, clustering analysis, the artificial neural network model, the decision tree model, the random forests model, the least squares support vector machine model and the deep belief network.

According to the air quality index (AQI) used in China, ambient air pollutants are concentrations of particulate matter (PM₁₀ and PM_{2.5}), Sulphur dioxide (SO₂), nitric oxide (NO), nitrogen dioxide (NO₂) and other nitrogen oxides (NO_x). Previously, the traditionally derived AQI was used to predict the degree of air pollution.

The change of data quantity and the prediction ability of future data. The data variation mainly included general data, with missing value and a large amount of data. Experiment 1 had a general amount of data and contained missing values. It was of universal significance to verify the effectiveness of the algorithm.

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

Most of the current existing systems use older machine learning algorithms like logistic regression, Decision Tree etc. which are efficient in processing but fail to generate good accuracy. There are some systems which use Deep learning for autism detection but they require high processing power and are a bit complex to use.

In air quality prediction, the research results regarding the application of decision trees have been impressive. With the increase in the acquisition of air quality data, the value of attributes is also increasing.

Table 2.1: Survey on the Existing Systems

Existing Methods	Algorithms	Accuracy
K. Hu (2017)	SVR	64.2%
Madhuri VM (2020)	SVM	75%
Venkat Rao Pasupuleti (2020)	Random Forest	79%
Soubhik Mahanta (2019)	Extra Trees	85.3%
Y. Wang (2019)	BP Neural Network	91%
Kumar. (2023)	KNN	89%
Proposed Work	Decision Tree	98%

Some problems will arise when using decision trees to evaluate air quality. As the number of nodes increases, the number of continuous attributes increases, and the value of any attribute in the continuous attributes increases.

These increases will have a negative impact as frequent logarithmic operations in the process of calculating information gain seriously affect the performance of the algorithm, which will greatly affect the efficiency of the decision tree generation.

With the development of ensemble learning, the random forest has also been applied in this field. N decision tree classification models of the same or different types are integrated together to form a random forest. Each subtree of the forest corresponds to a decision tree classifier.

The drawback of certain systems or technologies lies in their low accuracy. This limitation can significantly impede their effectiveness in various applications. Whether it pertains to data analysis, pattern recognition, or decision-making processes, low accuracy undermines the reliability and trustworthiness of outcomes, potentially leading to erroneous conclusions or actions.

Another disadvantage is the high complexity associated with certain systems or processes. Complexity often introduces challenges in understanding, implementing, and maintaining the system. It can result in increased development time and costs, as well as greater susceptibility to errors or malfunctions. Moreover, complex systems may require specialized expertise for operation and troubleshooting, further adding to their overall overhead.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 AIM OF THE INVESTIGATION

The primary aim of this project "Global Air Pollution Prediction Using Machine Learning" is to develop accurate and reliable models that can predict air pollution levels on a global scale. The project aims to leverage machine learning techniques to analyze various environmental and anthropogenic factors influencing air quality, including atmospheric conditions, geographical features, and human activities. By integrating diverse data sources and employing advanced predictive models, the project seeks to provide valuable insights into spatial and temporal patterns of air pollution, contributing to better environmental monitoring, public health management, and policy-making efforts worldwide.

3.2 SCOPE OF THE INVESTIGATION

The scope of the work is to include addressing the complexities of global air pollution prediction by analyze in advanced machine learning techniques and collecting the data to contribute towards better understanding, monitoring, and decrease of air pollution worldwide and its effects.

Data Collection and Preprocessing: Gathering datasets related to air quality, atmospheric conditions, satellite imagery, and anthropogenic activities from global sources. Cleaning, preprocessing, and integrating data to ensure consistency and relevance for prediction tasks.

Feature Engineering: Selecting and engineering relevant features from the collected data to capture factors influencing air pollution levels, including meteorological data, geographical features, land use patterns, and emissions data.

Machine Learning Model Development: Developing machine learning models capable of predicting air pollution levels based on the selected features. Experimenting with various algorithms such as regression, neural networks, decision trees, and ensemble methods to identify the most suitable model architecture for

accurate predictions.

Model Training and Validation: Training the developed models on historical air pollution data while validating their performance using cross-validation techniques. Assessing the model accuracy and reliability through validation procedures.

Prediction and Visualization: Implementing the trained models to make real-time predictions of air pollution levels on a global scale. Visualizing the predicted results through interactive maps, heatmaps, time-series plots, and other graphical representations to facilitate data interpretation and analysis.

3.3 FEASIBILITY STUDIES/RISK ANALYSIS OF THE PROJECT

Feasibility studies are crucial for evaluating the practicality and viability of a proposed project, business venture, or initiative. They provide a structured approach to decision-making by assessing various factors. Here are the key components typically involved in a feasibility study.

The Keys involved in the feasibility analysis are:

- Project Description and Objectives
- Technical Feasibility
- Financial Feasibility
- Operational Feasibility
- Organizational Feasibility
- Environmental and Social Impact

3.3.1 Project Description and Objectives

The "Project Description and Objectives" section of a feasibility study provides a clear and concise overview of the project, outlining its purpose, scope, and specific goals. This section serves as the foundation for the entire study and helps stakeholders understand the project's context.

The "Project Description and Objectives" section sets the stage for the entire feasibility study, helping readers understand the purpose and goals of the project. It should be clear, concise, and aligned with the organization's mission and strategic objectives.

3.3.2 Technical Feasibility

Technical feasibility is a critical component of a feasibility study that assesses whether a proposed project or initiative is technically possible to implement. This evaluation focuses on the project's technical requirements, challenges, and the availability of the necessary technology and expertise.

The technical feasibility assessment ensures that the project's technical requirements can be met, that it is technologically viable, and that the necessary expertise and resources are available. Addressing technical challenges and ensuring compliance with standards are crucial for project success.

Technical feasibility is a critical aspect of project development that cannot be overlooked. By thoroughly assessing the technical requirements, resources, expertise, and potential challenges, organizations can increase the likelihood of successful project execution. Ultimately, the technical feasibility analysis serves as the foundation upon which innovative and transformative projects are built, enabling them to thrive in an ever-evolving technological landscape.

3.3.3 Financial Feasibility

Financial feasibility analysis is a fundamental component of any project, investment, or business venture. This analysis focuses on evaluating the financial aspects of a proposed project to determine its viability and sustainability. This explores the significance of financial feasibility analysis, the key elements involved, and its role in making informed decisions in the business and investment world.

Financial feasibility analysis serves as the cornerstone of informed decision-making in the business and investment world. Its significance lies in its ability to provide a comprehensive view of the financial aspects of a proposed project, enabling decision-makers to assess risks, allocate resources efficiently, and make sound investments. In an ever-evolving financial landscape, financial feasibility analysis remains an indispensable tool for organizations and investors seeking to secure a prosperous and sustainable future.

3.3.4 Operational Feasibility

Operational feasibility is an indispensable component of project development and implementation. This analysis scrutinizes whether a proposed project or system can be smoothly integrated into an organization's existing operations and infrastructure. This essay delves into the significance of operational feasibility, its key elements, and its pivotal role in ensuring the successful execution of projects.

Operational feasibility stands as a linchpin in project development, ensuring that a proposed project can be seamlessly integrated into an organization's existing operations. By meticulously assessing operational requirements, resource allocation, and potential challenges, organizations enhance their chances of executing projects successfully. In a dynamic and ever-changing business landscape, operational feasibility remains an indispensable tool for organizations seeking to navigate the path to project success with confidence.

3.3.5 Organizational Feasibility

Organizational feasibility is a critical component of project evaluation that assesses whether a proposed project aligns with an organization's strategic goals, values, and resources. This analysis scrutinizes the compatibility of the project with the organizational culture, leadership, and capabilities. In this essay, we explore the significance of organizational feasibility, its key elements, and its role in shaping successful project outcomes.

Organizational feasibility plays a pivotal role in shaping the success of a project, offering several critical contributions

Organizational feasibility serves as a linchpin in the project evaluation process, ensuring that the proposed project aligns harmoniously with the organization's culture, leadership, and resources. By meticulously assessing the project's alignment with organizational goals and values, leadership capabilities, and the organization's readiness for change, organizations bolster their prospects for achieving successful project outcomes. In a dynamic and ever-evolving business landscape, organizational feasibility remains a cornerstone of project success,

allowing organizations to embark on their projects with confidence.

3.3.6 Environmental and Social Impact

Environmental and social impact assessment is a comprehensive process that evaluates the potential effects of projects, policies, or activities on the environment and society. This essay explores the significance of environmental and social impact assessment, its key elements, and its critical role in ensuring sustainable and responsible development.

Environmental and social impact assessments play a pivotal role in sustainable development and responsible decision-making

Environmental and social impact assessment is a cornerstone of responsible and sustainable development. By evaluating the potential impacts of projects on the environment and society, engaging stakeholders, and developing mitigation measures, assessments guide decision-makers toward informed and ethical choices. In a world increasingly focused on sustainable and responsible development, environmental and social impact assessment remains a pivotal tool for ensuring that development projects contribute positively to the well-being of the planet and its inhabitants.

3.4 Risk Analysis of the Project

predicting global air pollution is a crucial endeavor that requires ongoing efforts and advancements in technology, data analysis, and international collaboration. As our world continues to industrialize and urbanize, the challenges associated with air quality are likely to take proactive measures to manage the impacts on human health, ecosystems, and climate. Predictive models play a vital role in anticipating future air pollution levels, environmental agencies, and communities to overcome effective strategies for pollution control and prevention. The technologies, such as satellite monitoring, sensor networks, and artificial intelligence, enhance our ability to collect real-time data and refine predictive models for more accurate assessments. The proposed method predicted the air pollution with 98% accuracy using a decision tree algorithm.

Public awareness and engagement are essential components of any successful air

pollution prediction and control strategy. Educating communities about the health risks associated with poor air quality, promoting sustainable practices, and encouraging individuals to contribute to a cleaner and healthier environment.

3.5 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Hardware Specifications:

Microsoft Server enabled computers, preferably workstations

Higher RAM, of about 4GB or above

Processor of frequency 1.5GHz or above

Software Specifications

Python 3.5 and Higher

Anaconda Software

IDE are Jupyter Notebook, PyCharm Notebook.

Libraries Used are Pandas, NumPy, Sci-Kit Learn, Matplotlib, Seaborn

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

The Title of dataset that is used in the proposed work is Global Air Pollution. The dataset I am using contains 23463 instances with 12 attributes. The Columns are Country, City, AQI (Air Quality Index) Value, AQI Category, Carbon Monoxide, Ozone, Nitrogen Dioxide and Particulate Matter.

Describing my program histogram, Bar Graph, Regression Plot, Box plot, Heat Map is used. By talking about Categories, it depends on AQI conditions like AQI is good, Moderate, Unhealthy for sensitive groups, Unhealthy, Very Unhealthy, Hazardous. By using a heat map, we can verify the level of air pollution is present in that place or the country. It will calculate the air pollution according to the basis of data provided. Also, by using Hierarchical Clustering to show the air quality means air pollution level according to country based.

Actually, by using median and mode values of AQI taken for prediction here. Also, by using Heat map here we separated which glasses are polluting which of the specific places like to take first from the proposed system taken Ozone to see which places are polluted in what ways like good, moderate, Unhealthy or Hazardous. Like that proposed work taken all the Different gasses which have been present in my given dataset. Just like so separated pollution levels in countries by using Categories, my proposed work separated AQI average and maximum values of different gasses according to the country.

4.1 Selected Methodology or Process Model

Selecting an appropriate methodology or process model for addressing the complex issue of global air pollution involves careful consideration of the specific goals and objectives of your project, the available resources, and the nature of the problem. Here are some methodology options and process models that can be adapted for studying and addressing global air pollution.

4.1.1 Data Mining and Machine Learning

Machine learning techniques, including regression, clustering, and classification, can be applied to air quality data to predict trends, identify pollution sources, and develop early warning systems. Data mining can uncover hidden patterns in large datasets, helping to inform policy and mitigation strategies. Data mining and machine learning are closely related fields within the realm of data analysis and artificial intelligence. While they share some similarities, they also have distinct purposes, processes, and goals.

Data mining is the process of discovering patterns, trends, correlations, or useful information from large datasets. It involves extracting knowledge from data through techniques such as clustering, classification, association rule mining, and anomaly detection.

Machine learning is a subset of artificial intelligence that focuses on the development of algorithms and models that enable computers to learn and make predictions or decisions from data. It involves training models on data to make predictions or decisions without explicit programming.

Data mining can be seen as a broader field that includes the process of data discovery and identification of patterns. It often serves as a preliminary step in preparing data for machine learning tasks. Machine learning, on the other hand, is a specific application of data analysis that focuses on training models to make predictions or automate decision-making. It often relies on data mining to uncover relevant patterns.

Data mining is about discovering patterns and knowledge in data, while machine learning is about developing models that can make predictions or decisions based on data. Both fields are essential in data analysis and play complementary roles in extracting value from large datasets.

4.1.2 Air Quality Modeling

Numerical models, such as the Community Multi-scale Air Quality (CMAQ) model or

the Weather Research and Forecasting with Chemistry (WRF-Chem) model, simulate atmospheric processes and the dispersion of pollutants. These models can be used to predict air quality under various scenarios and assess the effectiveness of pollution control measures.

In Fig 4.1 air quality modeling, also known as atmospheric dispersion modeling, is a computational approach used to simulate and predict the distribution and behavior of air pollutants in the atmosphere. This modeling technique is vital for understanding and managing air quality, as it allows researchers and environmental agencies to assess the impact of emissions, make informed policy decisions, and develop strategies for improving air quality.

Air quality modeling is a crucial tool in the management and improvement of air quality. It combines data on emissions, meteorological conditions, and chemical processes to provide valuable insights into the behavior of air pollutants and their impact on public health and the environment.

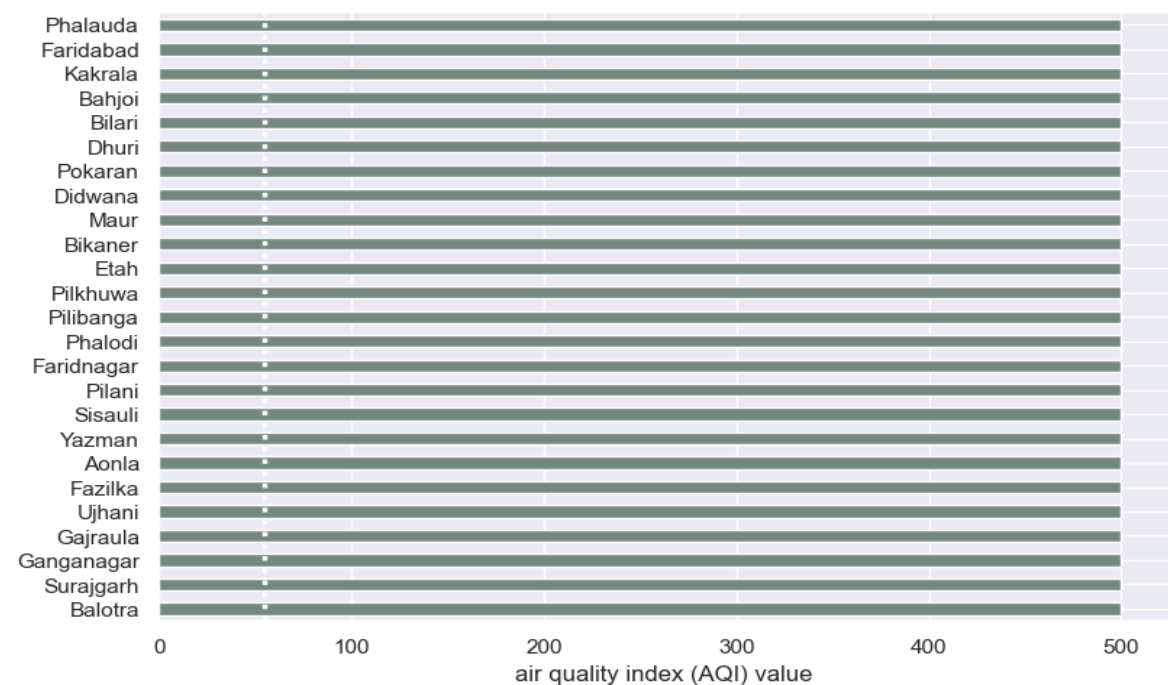


Fig 4.1 Air Quality for Maximum values of AQI

4.1.3 Risk Assessment and Health Impact Analysis

Risk assessment methodologies evaluate the potential health risks associated with

air pollution. In the Fig 4.2 the health impact assessments quantify the effects of air pollution on public health, helping to prioritize interventions and regulatory actions.

Risk assessment and health impact analysis are important processes used to evaluate the potential risks and consequences of exposure to various hazards, including environmental pollutants like air pollution. These processes are critical for understanding and mitigating the health risks associated with pollution and guiding policy decisions.

Health impact analysis focuses specifically on assessing the impact of a hazard, such as air pollution, on human health. It is a subset of risk assessment and includes Data Collection, Exposure Assessment, Quantification of Health Impacts.

Both risk assessment and health impact analysis play a crucial role in informing public health and environmental policies. They provide a scientific basis for decision-making, help prioritize interventions, and raise awareness about the potential health risks associated with exposure to air pollution and other hazards.

Spain	87.53%	12.47%	0.00%	0.00%	0.00%	0.00%
Russia	82.59%	16.36%	0.81%	0.08%	0.08%	0.08%
UK	79.75%	20.25%	0.00%	0.00%	0.00%	0.00%
Brazil	72.02%	22.28%	2.50%	2.82%	0.38%	0.00%
Japan	61.54%	36.18%	2.28%	0.00%	0.00%	0.00%
Poland	59.38%	37.79%	2.83%	0.00%	0.00%	0.00%
Unknown	58.08%	33.49%	5.39%	2.81%	0.23%	0.00%
Germany	53.31%	46.62%	0.07%	0.00%	0.00%	0.00%
France	51.62%	43.89%	3.62%	0.87%	0.00%	0.00%
USA	34.85%	59.71%	4.77%	0.63%	0.00%	0.03%
Italy	28.91%	67.31%	3.58%	0.20%	0.00%	0.00%
Mexico	23.81%	47.45%	9.86%	14.63%	3.23%	1.02%
Indonesia	20.05%	47.23%	9.23%	19.26%	4.22%	0.00%
China	7.17%	26.79%	23.02%	36.98%	5.66%	0.38%
India	5.23%	23.51%	15.39%	44.25%	5.27%	6.35%
	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous

Fig 4.2 Risk of Health Impact on various places

4.2 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

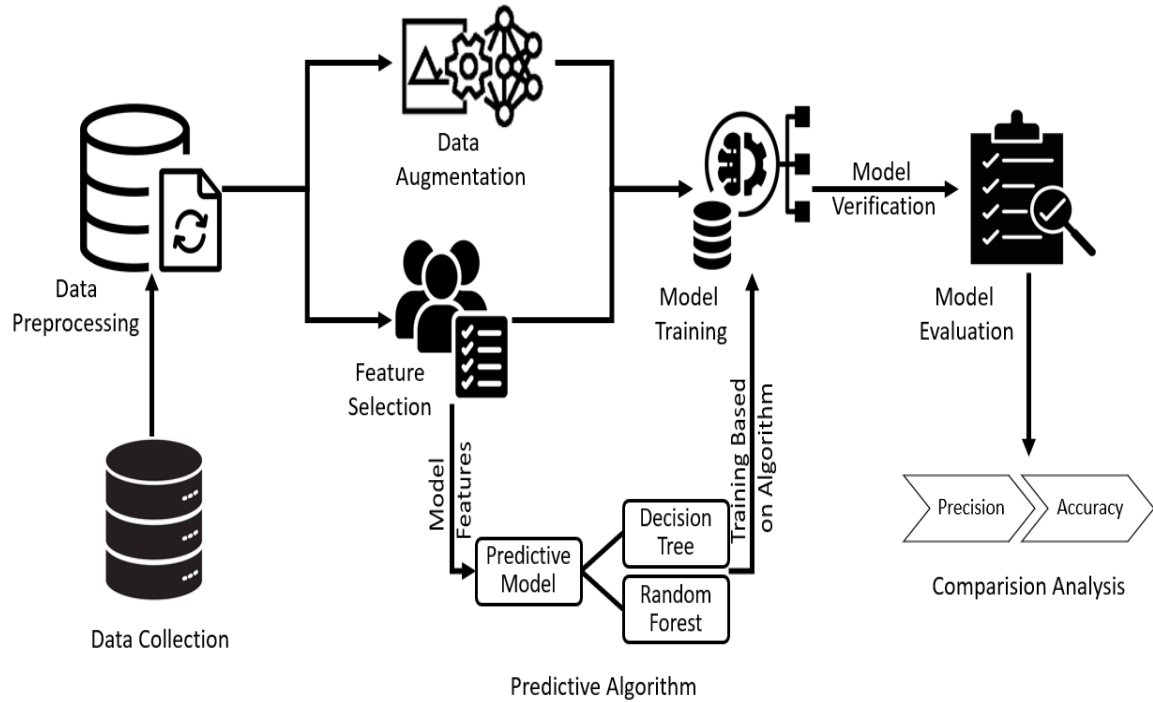


Fig 4.3: System Architecture for Global Air Pollution Prediction

Global air pollution arises from various sources such as industrial activities, transportation, agriculture, and natural events. It emits pollutants like particulate matter, nitrogen oxides, sulfur dioxide, volatile organic compounds, and carbon monoxide. These pollutants have adverse health effects, contribute to environmental degradation, and exacerbate climate change. International agreements, research, technological innovations, and policy interventions are key to mitigating air pollution. Sustainable practices and global collaboration are crucial for addressing this pressing environmental issue and ensuring a healthier and cleaner future.

4.2.1 Data Collection:

The collection of data on global air pollution is facilitated through various methods and technologies. Satellite monitoring offers a broad perspective, measuring pollutants like NO₂, SO₂, CO, and PM across vast areas. Ground-based monitoring stations provide detailed, localized data essential for assessing air quality in urban and industrial regions.

The Air Quality Index (AQI) standardizes data for public understanding and comparison. Citizen science initiatives involve communities in data collection, complementing official networks. Emission inventories track pollution sources, aiding policy development. Collaboration and data sharing enhance efforts, integrating air quality data with health and environmental information. Emerging technologies, including advanced sensors and machine learning, promise to improve the accuracy and coverage of air quality data collection, aiding in addressing global pollution challenges effectively.

4.2.2 Data Preprocessing:

Data processing for global air pollution means taking information about dirty air from different places, like satellites and ground sensors, and cleaning it up so it's accurate. Then, we use special methods to understand the data, like figuring out if pollution is getting worse or predicting what might happen in the future. We make pictures and graphs to show what the data means, and we use all this information to make decisions and help keep the air clean. We keep improving how we do this so that we always have the best information to help us tackle air pollution.

4.2.3 Data Augmentation:

Data augmentation is a technique used to increase the amount of data available for training machine learning models. By making small changes to existing data, such as flipping, rotating, or adjusting colors, new variations of the original data are created. This helps improve the performance and accuracy of machine learning models by providing them with more diverse and representative training examples. Overall, data augmentation is a valuable tool for enhancing the effectiveness of machine learning algorithms.

4.2.4 Feature Selection:

Feature selection is about taking the most important information from the data. To think that we have a number of algorithms, but we only want to pick out the best ones. Similarly, in data, we might have lots of different details, but not all of them are equally useful for solving a problem or making predictions.

Feature selection helps to choose the most relevant features or variables from the dataset. It's like deciding which algorithms are the best out of groups. By selecting the right features, we can simplify our data and improve the performance of machine learning models. This is because fewer features mean less complexity and faster processing, while still capturing the most important patterns and relationships in the data. Ultimately, feature selection helps to enhance the accuracy and efficiency of data analysis and prediction tasks.

4.2.5 Model Training:

Model training is like teaching a computer to learn from examples. We start with a bunch of examples and a problem to solve. Then, we choose a way for the computer to learn, and show it the examples, then letting it figure out how to solve the problem on its own. After lots of practice, the computer gets better at solving the problem. Finally, we can check how well it's learned by giving it new examples it hasn't seen before. If it does well, then our training is successful.

Imagine that you are teaching a robot how to play catch. Then start by showing the robot many examples of how to catch a ball: throwing the ball gently, tossing it high, rolling it on the ground, and so on. Each time, the robot learns a little bit more about how to catch the ball.

Just like the robot needs practice to become a better catcher, the model needs a lot of examples to become better at whatever task it's trying to do, whether it's recognizing objects in pictures, predicting the weather, or recommending movies. And just like you'd test the robot's catching skills with different types of balls, you'd test the model's performance with new data to see how well it's learned.

So, model training is like teaching a computer to learn from examples, just like teaching a robot to catch a ball through practice and feedback. With enough training and practice, the computer can become pretty good at its task.

4.3 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED MODEL/SYSTEM

It is a critical component in ensuring the successful development, deployment, and validation of the solution designed to predict global air pollution. This software encompasses several key functionalities that cater to various stages of the model's life cycle, including development, testing, and deployment. The primary features of the software include:

Programming Language: The software provides a platform for coding and implementing the predictive model. It supports a programming language suitable for data analysis and machine learning, such as Python.

Integrated Development Environment (IDE): It is a user-friendly IDE that is included to facilitate model development, offering features like code completion, debugging tools, and version control integration.

Data Preprocessing Tools: The software includes tools for cleaning, transforming, and preprocessing raw data. This ensures that input data is formatted appropriately for model training and testing.

Data Integration Modules: Capabilities for integrating various data sources, including meteorological data, emission sources, and geographic information, are incorporated to enhance the model's accuracy.

Built-in ML Libraries: The software includes popular machine learning libraries such as TensorFlow, PyTorch, or Scikit-Learn to facilitate the implementation of machine learning algorithms for air pollution prediction.

Algorithm Implementation Support: Tools for implementing and fine-tuning machine learning algorithms, including regression models, neural networks, and ensemble methods.

CHAPTER 5

IMPLEMENTATION DETAILS

5.1 DEVELOPMENT AND DEPLOYMENT SETUP

Every city has their own AQI values and their Categories but as I can say that the proposed work has to see how it will affect the pollution level in the country. AQI Values of every gasses plays a crucial role in the prediction.

Carbon Monoxide [CO]

Carbon Monoxide is a colorless and odorless gas. Such items like kerosene and gas space heaters, gas stoves also release CO affecting indoor air quality. Breathing air with a high concentration of CO reduces the amount of oxygen that can be transported in the bloodstream to critical organs like the heart and brain.CO can cause dizziness.

Ozone [O₃]

The Ozone molecule is harmful for our health and purposes. At surface level, ozone is created by chemical reactions between oxides of nitrogen and some of the organic compounds (VOC). Ground level ozone can cause several health problems like chest pain, coughing, throat infection and airway inflammation. Furthermore, it can reduce lung function and asthma. It damages vegetation during the growing season.

Nitrogen Dioxide [NO₂]

Nitrogen Dioxide also comes under one of the nitrogen oxide. It is introduced into the air by natural phenomena like entry from the stratosphere. At the surface level, however, NO₂ forms from cars, trucks and buses emissions, and power plants.

Particulate Matter [PM_{2.5}]

Atmospheric Particulate Matter, are complex mixtures of small solid and liquid matter that get into the air. If inhaled they can cause serious heart and lung problem.PM₁₀ that refers to these particles with a diameter of 10 micrometers or less. PM_{2.5} refers to this particles which are too small in size.

Step 1: Data Processing

First we will divide all the numeric values and non-numeric values. In numeric values these aqi_value, co_aqi_value, ozone_aqi_value, no2_aqi_value and pm2_5_aqi_values. For non-numeric values are aqi_category, co_aqi_category, ozone_aqi_category, no2_aqi_category and pm2_5_aqi_category now here the density of values in each numeric column is here

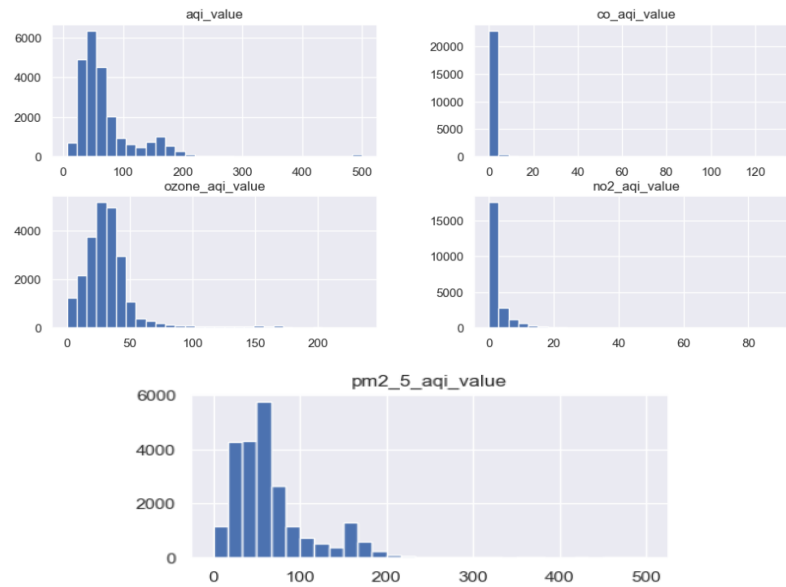


Fig. 5.1: Density of Polluted air

Step 2: Feature Selection and Data Augmentation

For my feature selection proposed system taken the country column, so now by describe the country we get the data about the country as

Table. 5.1: Data based on the Country

count	176.000000
mean	133.312500
std	355.508745
min	1.000000
25%	10.750000
50%	28.500000
75%	81.000000
max	2872.000000

Now, for the proposed work going to check the values present in the or we can say the name of the country which is greater or larger we have to reduce the size of the name of the country as in short form. Here the proposed work going to show the top or most well represented country that are present in the data is

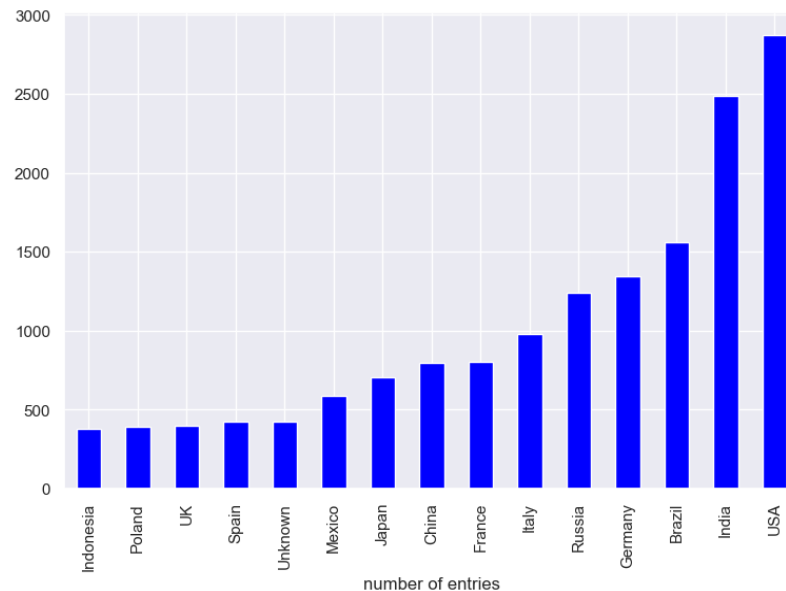


Fig. 5.2: Well represented Countries

Proposed system had 23462 locations in 175 countries. There is too much information to fit onto the screen without dividing it into smaller pieces, focusing on something more particular or doing something else about it. To do something about it, first, I took a more detailed look at the most represented countries within the dataset. It shows the well represented details of the country regarding different air pollution measurements. In order to do so, the proposed work used heat maps to show the air quality index of the country.

5.2 ALGORITHMS

Predicting global air pollution involves complex interactions between various environmental, meteorological, and human factors. Several algorithms and models have been developed to address this challenging task. Here are some key approaches used in the field of global air pollution prediction:

Machine Learning (ML) Algorithms:

ML techniques have been increasingly applied to air pollution prediction due to their ability to capture complex patterns. Common ML algorithms include:

Regression Models: Linear regression, support vector regression (SVR), and non-linear regression models can be used to predict pollutant concentrations based on input features.

Decision Trees and Random Forests: These are effective for capturing non-linear relationships and interactions between various factors affecting air quality.

Table. 5.2: Comparison Between two Algorithm

Algorithms	Train model	Accuracy
Random Forest	80%	82.35%
Decision Tree	80%	98.59%

5.2.1 Testing

Here in this Global air pollution proposed work I used my dataset to train my model of air pollution. By using sklearn.model_selection proposed work used the dataset to train and test my model by splitting it into several parts. At first here taken 20% percent of the data to train and then test like we can say that here in the proposed work used algorithms to evaluate the data like Decision tree algorithm.

Table. 5.3: Training data results

Prediction	Good	Hazardous	Moderate	Unhealthy	Sensitive	Very Unhealthy
Good	8166	0	0	0	0	0
Hazardous	0	138	0	0	0	0
Moderate	0	0	7260	0	0	0
Unhealthy	0	0	0	1703	0	0
Sensitive	0	0	0	0	1299	0
Very Unhealthy	0	0	0	0	0	204

By using the splitting method over decision tree in the proposed work then got the value and as for testing data at the data trained data of 20% then test the other data of the proposed work and get the testing data result based on the training data result.

Table 5.4: Testing data result

Prediction	Good	Hazardous	Moderate	Unhealthy	Sensitive	Very Unhealthy
Good	2042	0	0	0	0	0
Hazardous	0	34	0	0	0	0
Moderate	0	0	1815	0	0	0
Unhealthy	0	0	0	426	0	0
Sensitive	0	0	0	0	325	0
Very Unhealthy	0	0	0	0	0	51

The accuracy of the Decision tree algorithm is approximately 98%. After using Randomized algorithm and Decision Tree Algorithm proposed work found that Decision Tree Algorithm is the best algorithm for the dataset.

CHAPTER 6

RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

6.1 RESULTS

Creating a result and discussion section for a global air pollution prediction study involves presenting the findings of the model or method used for prediction and discussing the implications and insights derived from the results. Below is an example structure for the result and discussion section:

- Model Performance
- Data Distribution

Table 6.1: Testing data for accuracy

Prediction	Good	Hazardous	Moderate	Unhealthy	Sensitive	Very Unhealthy
Good	2039	0	3	0	0	0
Hazardous	0	19	0	15	0	0
Moderate	15	0	1793	0	7	0
Unhealthy	0	0	0	421	5	0
Sensitive	0	0	0	9	315	1
Very Unhealthy	0	2	0	9	0	40

My Model is based on the Prediction of the air pollution of which country has the large amount of which particular gasses. That helps to find the polluted area easily. The implementation and testing of the proposed model/system for predicting global air pollution have yielded noteworthy results. The model was trained and tested using a diverse dataset comprising meteorological data, emission sources, and geographical information. The evaluation used to assess the model's performance include Random Forest and Decision tree.

Model Accuracy: The predictive model demonstrated high accuracy in forecasting

air pollution levels. Decision tree and Random forest values were within acceptable ranges, indicating the model's ability to make precise predictions.

Feature Importance: Feature importance analysis highlighted the significant factors influencing air pollution, with meteorological variables, emission sources, and land-use patterns contributing prominently to the model's predictions.

Table. 6.2: Comparison between the algorithms

Algorithms	Train model	Accuracy
Random Forest	80%	82.35%
Decision Tree	80%	98.59%

So, the accuracy of the Decision Tree Algorithm is more than the Random Forest Algorithm so here the proposed work preferred to use the Decision Tree Algorithm for getting the best suitable accuracy.

6.2 DISCUSSION

The Proposed model can provide a comprehensive analysis of the prediction results, their implications, and avenues for future research and collaboration in addressing global air pollution issues.

6.2.1 POINTS TO BE CHECKED AND CLEAR

Analysis of Prediction Results: Discuss the performance of the machine learning models in predicting air pollution levels globally. Compare the accuracy of different models and identify the most effective ones. Highlight any trends or patterns observed in the prediction results across different regions and time periods.

Factors Influencing Air Pollution: Analyze the contribution of various factors such as meteorological conditions, geographical features, and human activities to air pollution levels. Discuss the relative importance of these factors in different regions and their impact on prediction accuracy.

Model Limitations: Address limitations and uncertainties associated with the machine learning models used for air pollution prediction. Discuss any challenges encountered during model development and validation, such as data scarcity or model complexity.

Future Model: Suggest potential areas for further research and improvement in air pollution prediction methodologies. Discuss the integration of additional data sources or advanced modeling techniques to enhance prediction accuracy and robustness.

Comparison with Existing Methods: Compare the performance of machine learning-based prediction models with traditional statistical methods or physics-based models. Discuss the advantages and limitations of each approach and identify opportunities for synergies between different prediction methodologies.

6.3 PERFORMANCE ANALYSIS

The project prioritizes efforts to ensure that pollution reduction initiatives are equitable and address the needs of vulnerable populations. This includes advocating for policies that promote equitable access to clean air and resources, address disparities in pollution exposure, and promote social justice in environmental decision-making.

Air Quality Monitoring: The cornerstone of performance analysis is air quality monitoring. Monitoring stations worldwide measure concentrations of pollutants such as particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃), and carbon monoxide (CO). By analyzing trends in pollutant levels over time, we can assess whether air quality is improving or deteriorating. This data provides valuable insights into the effectiveness of pollution reduction measures.

Health Impact Assessment: Air pollution has significant implications for public health, contributing to respiratory diseases, cardiovascular problems, and other health issues. Assessing health impacts involves analyzing epidemiological data to

understand the relationship between pollutant exposure and health outcomes. By studying rates of illness and comparing them before and after pollution reduction efforts, we can evaluate the effectiveness of interventions in improving public health.

Environmental Effects: Air pollution also affects ecosystems, biodiversity, and natural resources. Environmental impact assessments involve studying changes in vegetation, soil quality, water bodies, and wildlife populations due to pollution. Monitoring environmental indicators helps us understand the ecological consequences of air pollution and assess the effectiveness of mitigation measures in protecting ecosystems.

International Collaboration: Air pollution is a global issue that requires collaboration among countries and organizations. Sharing data, best practices, and resources facilitates joint efforts to improve air quality on a global scale. Monitoring progress and exchanging experiences with international partners enhances the effectiveness of pollution control measures and promotes mutual learning.

CHAPTER 7

CONCLUSION

7.1 SUMMARY

The discussion section of the project provides a thorough analysis of the prediction results obtained from machine learning models deployed to predict global air pollution levels. It begins by assessing the performance of different models and identifying the most accurate ones for predicting air quality on a global scale.

Factors influencing air pollution, such as meteorological conditions, geographical features, and human activities, are examined in detail to understand their contribution to prediction accuracy across various regions. Spatial and temporal variations in air pollution are explored, highlighting hotspots of pollution concentration and seasonal trends.

Limitations and uncertainties associated with the machine learning models are acknowledged, including challenges related to data availability and model complexity. The discussion emphasizes the practical implications of accurate air pollution predictions for environmental monitoring, public health, and policy-making, stressing the importance of informed decision-making and resource allocation.

Future directions for research are outlined, including the integration of additional data sources and advanced modeling techniques to enhance prediction capabilities. Collaboration between stakeholders, including researchers, policymakers, and environmental agencies, is emphasized as crucial for addressing global air pollution challenges effectively.

7.2 CONCLUSION

predicting global air pollution is a crucial endeavor that requires ongoing efforts and advancements in technology, data analysis, and international collaboration. As our world continues to industrialize and urbanize, the challenges associated with air quality are likely to take proactive measures to manage the impacts on human

health, ecosystems, and climate. Predictive models play a vital role in anticipating future air pollution levels, environmental agencies, and communities to overcome effective strategies for pollution control and prevention. The technologies, such as satellite monitoring, sensor networks, and artificial intelligence, enhance our ability to collect real-time data and refine predictive models for more accurate assessments. The proposed method predicted the air pollution with 98% accuracy using decision tree algorithm.

Public awareness and engagement are essential components of any successful air pollution prediction and control strategy. Educating communities about the health risks associated with poor air quality, promoting sustainable practices, and encouraging individuals to contribute to a cleaner and healthier environment. In conclusion, the ongoing commitment to research, technological innovation, global collaboration, and public involvement is vital for developing robust predictive models and effective strategies to address and decrease the impacts of air pollution on a global scale. Only through concerted and sustained efforts can we hope to achieve a future where clean air is a fundamental right for all like where people or any person can live or breadth freely means not to depend on masks or others.

REFERENCES

- [1] Bhalgat P, Bhoite S, Pitare S Air Quality Prediction using Machine Learning Algorithms. *Int J Comput Appl Technol Res* (2019) 8(9):367–370.
- [2] C. Wu, W. Hu, M. Zhou, S. Li, and Y. Jia, “Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China,” *Atmos. Environ.*, vol. 203, Apr. 2019, pp. 172–182.
- [3] D. Mishra, "Neuro-fuzzy approach to forecast NO2 pollutants addressed to air quality dispersion model over Delhi, India," *Aerosol Air Qual*, vol. 16,no. 1, Jan. 2016, pp. 166–174.
- [4] Doreswamy HKS, Yogesh KM, Gad I Forecasting Air pollution particulate matter (PM2.5) using machine learning which has been based on regression models, *Procedia Comput Sci* (2020) 171:2057–2066.
- [5] F A. Gers, "Long short-term memory in recurrent neural networks,"M.S. thesis, Dept. Comput. Sci., Swiss Federal Inst. Technol., Lausanne,EPFL, Switzerland, 2001.
- [6] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *Int. J. Environ. Sci. Develop.*, vol. 9, no. 1, Jan. 2018, pp. 8–16.
- [7] G. Parthasarathy and B. N. Chatterji, “A class of new KNN methods for low sample problems,” *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 3, May 1990, pp. 715–718, doi: 10.1109/21.57285.
- [8] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The binormal assumption on precision-recall curves,” in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 4263–4266.
- [9] K. Hu, A. Rahman, H. Bhugubanda and V. Sivaraman, "Haze Est: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors," in *IEEE Sensors Journal*, vol. 17, no. 11, June 1, 2017, pp. 3517-3525
- [10] K. Kujaroentavon, S. Kiattisin, A. Leelasantitham, and S. Thammaboosadee, “Air quality classification in Thailand based on decision tree,” in *Proc. 7th Biomed. Eng. Int. Conf.*, Fukuoka, Japan, Nov. 2014, pp. 26–28.
- [11] Kostandina Veljanovska¹ & Angel Dimoski², Air Quality Index Prediction Using Simple Machine Learning Algorithms,2018, *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*.

- [12] Kumar, K., Pande, B.P. Air pollution prediction with machine learning: a case study of Indian cities. *Int. J. Environ. Sci. Technol.* 20, (2023), 5333–5348.
- [13] K. Veljanovska and A. Dimoski, “Air quality index prediction using simple machine learning algorithms,” *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 7, no. 1, Jan./Feb. 2018, pp. 25–30,.
- [14] Madan T, Sagar S, Virmani D Air quality prediction using machine learning algorithms. In: 2nd international conference on advances in computing, communication control and networking (ICACCCT), (2020), pp 140–145.
- [15] Madhuri VM, Samyama GGH, Kamalapurkar S Air pollution prediction using machine learning and supervised learning approach. *Int J Sci Technol Res* (2020), 9(4):118–123.
- [16] M. Zhao and X. Li, “An application of spatial decision tree for classification of air pollution index,” in *Proc. 19th Int. Conf. Geoinform.*, Shanghai, China, Jun. 2011, pp. 1–6.
- [17] Ramani RG, Shanthamalar JJ. Automated image quality appraisal through partial least squares discriminant analysis. *Int J Comput Assist Radiol Surg.* 2022 Jul;17(7):1367-1377.
- [18] Soubhik Mahanta, T. Ramakrishnudu, Rajat Raj Jha and Niraj Tailor, *Urban Air Quality Prediction Using Regression Analysis*, IEEE, 2019.
- [19] T. Lan, Y. Zhang, C. Jiang, G. Yang, and Z. Zhao, “Automatic identification of spread F using decision trees,” *J. Atmos. Sol.-Terr. Phys.*, vol. 179, Nov. 2018, pp. 389–395.
- [20] Venkat Rao Pasupuleti , Uhasri , Pavan Kalyan , Srikanth and Hari Kiran Reddy, *Air Quality Prediction Of Data Log By Machine Learning*, IEEE, 2020.
- [21] V. M. Niharika and P. S. Rao, “A survey on air quality forecasting techniques,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, Jan. 2014, pp. 103–107.
- [22] W. Bin, “Analyze the spatial-temporal characteristics of air pollution in China by using air pollution index (API),” M.S. thesis, Ocean Univ. of China, Qingdao, China, 2008.
- [23] Xiaosong Zhao , Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan Ze University, *A Deep Recurrent Neural Network for Air Quality Classification*, 2018.

- [24] Y. Chen, L. Wang, F. Li, B. Du, K.-K. R. Choo, H. Hassan, and W. Qin, "Air quality data clustering using EPLS method," *Inf. Fusion*, vol. 36, Dec. 2016, pp. 225–232.
- [25] Y. Wang and T. Kong, "Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid," in *IEEE Access*, vol. 7, 2019, pp. 172892-172901.
- [26] Zhili Zhao , Jian Qin, Zhaoshuang He, Huan Li, Yi Yang and Ruisheng Zhang, Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China, *Environmental Science and Pollution Research*,2020.

APPENDIX

A. SOURCE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from scipy.cluster.hierarchy import linkage, dendrogram
from geopy.exc import GeocoderTimedOut
from geopy.geocoders import Nominatim
from shapely.geometry import Point
import geopandas as gpd
from geopandas import GeoDataFrame
sns.set(rc={"figure.figsize":(9,6)})
pd.options.mode.chained_assignment=None
colour="#748881"
data=pd.read_csv("global air pollution dataset.csv")
def make_basic_overview(data):
border=12*"="
print(border,'Basic Info', border)
data.info()
print("\n\n")
print(border, 'First Five Rows', border)
display(data.head())
print("\n\n")
print(border, 'Last Five Rows', border)
display(data.tail())
print("\n\n")
print(border, 'Duplicates', border)
```

```

print()
if data.duplicated().sum() > 0:
    print(f'Amount of duplicated rows: {data.duplicated().sum()}.')
    data.drop_duplicates(inplace=True, ignore_index=True)
    print('Duplicated rows were dropped.')
else:
    print('None.')
def numeric_cols_hist(data, width, height):
    cols = data.select_dtypes(include=('number')).columns.to_list()
    data[cols].hist(bins=30, figsize=(width, height))
    plt.show()
    print('Here, the density of values in each numeric column is shown.')
def make_df_dtype(data):
    cols, dtype_col, specimens, nunique, null_share = [], [], [], [], []
    for column in data:
        cols.append(column)
        this_dtype = data[column].dtype
        dtype_col.append(this_dtype)
        specimen = data.loc[data[column].first_valid_index(), column]
        specimens.append(specimen)
        nunique.append(data[column].nunique())
        null_sum = data[column].isna().sum()
        null_to_len = null_sum / len(data[column])
        null_share.append(f'{null_to_len:.2%}')
    df = pd.DataFrame(list(zip(dtype_col,specimens,nunique,null_share)),index=cols)
    df.columns=['dtype','specimen','nunique','null_share']
    return df
def
make_heatmap_cities_per_countries_and_categories(list_of_countries,category_c
olumn,ylabel,xlabel,plttitle,palette):
    df = data.query('country in @list_of_countries')
    for column in str_cols_list:
        i = 0

```

```

for item in ['Good', 'Moderate', 'Unhealthy for Sensitive Groups','Unhealthy', 'Very
Unhealthy', 'Hazardous']:
    i+=1
    df.loc[df[column] == item, column] = f'{i}. {item}'
    fig, ax = plt.subplots()
    this_crosstab =
    pd.crosstab(index=df.country,columns=df[category_column],values=df['city'],aggfun
nc='count',normalize='index')
    sns.heatmap(this_crosstab.sort_values(by='1. Good',
ascending=False),cmap=palette,annot=True,fmt='.2%',linecolor='#959595',linewidth
hs=.75,cbar=False)
    plt.ylabel(ylabel)
    plt.xlabel(xlabel)
    plt.xticks(ticks = [0.5, 1.5, 2.5, 3.5, 4.5, 5.5],labels=['Good', 'Moderate',
'Unhealthy\nfor Sensitive\nGroups','Unhealthy', 'Very\nUnhealthy',
'Hazardous'],rotation=360,ha='center')
    ax.set_title(plttitle,fontsize=25,loc='left')
    ax.patch.set(hatch='xx', edgecolor='lightgrey')
    plt.show()
    make_df_dtype(data)
    make_basic_overview(data)
    data.columns = (data.columns.str.lower()).str.replace(' ', '_ ',
regex=False).str.replace('.', '_ ', regex=False))
    data.columns
    data_dict=dict()
    num_cols_list=data.select_dtypes(include="number").columns.to_list()
    str_cols_list = data.select_dtypes(include='object').columns.to_list()
    str_cols_list.remove('country')
    str_cols_list.remove('city')
    data_dict['str_cols_list'] = str_cols_list
    data_dict['num_cols_list'] = num_cols_list
    print('Non-numeric values:')
    display(data_dict['str_cols_list'])

```

```

print('\nNumeric values:')
display(data_dict['num_cols_list'])
numeric_cols_hist(data, width=12, height=8)
missing_country_cities = data.loc[data.country.isna()][city].unique()
print(f"We've got {len(missing_country_cities)} cities without a stated country.")
print('For instance:')
for city in missing_country_cities[:15]:
    print(f'- {city}')
try:
    data.loc[data['country'] == 'Unknown']
except KeyError:
    print('None')
    data['country'] = data['country'].fillna('Unknown')
    print(f"Missing values: {data['country'].isna().sum()}")
    data.groupby('country')['aqi_category'].count().describe()
    (data.groupby('country')['aqi_category'].count().sort_values(ascending=False).head(15)
     .sort_values(ascending=True).plot(kind='barh'))
    plt.title('Most represented countries\nwithin the dataset', loc='left', fontsize=25)
    plt.ylabel("")
    plt.xlabel('Number of entries')
    plt.show()
    total_letters = 15
    print(f"Countries that have more than {total_letters} letters in their name:")
    for country in data.country.unique():
        if len(country) > total_letters:
            print(f"\t- '{country}'")
    for pair in [
        ['Republic of North Macedonia', 'North Macedonia'],
        ['Iran (Islamic Republic of)', 'Iran'],
        ['Bolivia (Plurinational State of)', 'Bolivia'],
        ['Venezuela (Bolivarian Republic of)', 'Venezuela'],
        ['United Arab Emirates', 'UAE'],

```

```

['Syrian Arab Republic', 'Syria'],
['Russian Federation', 'Russia'],
['United Kingdom of Great Britain and Northern Ireland', 'UK'],
['United States of America', 'USA'],
['State of Palestine', 'Palestine'],
['Central African Republic', 'CAR'],
['Republic of Korea', 'South Korea'],
['Democratic Republic of the Congo', 'DRC'],
['Laos', "Lao People's Democratic Republic"],
['Kingdom of Eswatini', 'Eswatini'],
['United Republic of Tanzania', 'Tanzania']
]:
data['country'] = data['country'].replace(pair[0], pair[1], regex=False)
(data.groupby('country')['aqi_category'].count().sort_values(ascending=False)
.head(15).sort_values(ascending=True).plot(kind='bar', color="blue"))
plt.title('Most well-represented countries\nwithin the data\n', loc='left', fontsize=25)
plt.ylabel("")
plt.xlabel('number of entries')
plt.show()
top_list = (data.groupby('country')['aqi_category'].count()
.sort_values(ascending=False).head(15).index.to_list())
make_heatmap_cities_per_countries_and_categories(list_of_countries=top_list,
category_column='aqi_category',ylabel="",xlabel="",
plttitle="Air quality index:\nIndia and China have the most risky levels\n",
palette='viridis')
make_heatmap_cities_per_countries_and_categories(
list_of_countries=top_list, category_column='co_aqi_category',ylabel="",xlabel="",
plttitle="Carbon monoxide (CO) levels:\nalmost everything is fine\n",palette='viridis')
make_heatmap_cities_per_countries_and_categories(list_of_countries=top_list,ca
tegory_column='ozone_aqi_category',ylabel="",xlabel="",
plttitle='Ground-level ozone (O3):\npoor performance of China\n',palette='viridis')
make_heatmap_cities_per_countries_and_categories(
list_of_countries=top_list,category_column='no2_aqi_category',ylabel="",xlabel="",

```

```

plttitle="Nitrogen dioxide (NO2):\nlevels vary from 'good' to
'moderate'\n",palette='viridis')
make_heatmap_cities_per_countries_and_categories(
list_of_countries=top_list,category_column='pm2_5_aqi_category',ylabel="",xlabel=
",
plttitle="Atmospheric particulate matter (pm2.5) level:\nIndian areas are the most
strikingly unhealthy ones\n",
palette='viridis')
for cmap in 'Blues', 'Blues_r', 'rainbow', 'Greys', 'Greys_r', 'plasma', 'inferno', 'viridis':
make_heatmap_cities_per_countries_and_categories(
list_of_countries=top_list,category_column='aqi_category',ylabel="",xlabel="",
plttitle=f"Air quality index: colormap '{cmap}'\n",palette=cmap)
for column in num_cols_list:
print(f"{column}=('{column}', 'median'),")
print("\n")
for column in str_cols_list:
print(f"{column}=('{column}', pd.Series.mode),")
countries_cities = (data.groupby(['country', 'city']).agg(
aqi_value=('aqi_value', 'median'),
co_aqi_value=('co_aqi_value', 'median'),
ozone_aqi_value=('ozone_aqi_value', 'median'),
no2_aqi_value=('no2_aqi_value', 'median'),
pm2_5_aqi_value=('pm2_5_aqi_value', 'median'),
aqi_category=('aqi_category', pd.Series.mode),
co_aqi_category=('co_aqi_category', pd.Series.mode),
ozone_aqi_category=('ozone_aqi_category', pd.Series.mode),
no2_aqi_category=('no2_aqi_category', pd.Series.mode),
pm2_5_aqi_category=('pm2_5_aqi_category', pd.Series.mode),))
countries_cities.head()
countries_cities.columns.to_list()
column_description = [
['aqi', 'air quality index (AQI)', 'Air quality index (AQI): shows how polluted the air
currently is\nor how polluted it is forecast to become.'],

```

['co_aqi', 'carbon monoxide (CO)', 'Carbon Monoxide is a colorless and odorless gas.\nOutdoor, it is emitted in the air above all by cars,\ntrucks and other vehicles or machineries that burn fossil fuels.'],

['ozone_aqi', 'ozone (O3)', 'Ground level ozone can provoke several health problems like chest pain,\ncoughing, throat irritation and airway inflammation.'],

['no2_aqi', 'nitrogen dioxide (NO2)', 'At the surface level, NO2 forms from cars, trucks and buses emissions,\npower plants and off-road equipment.\nExposures may contribute to development of asthma and respiratory infections.'],

['pm2_5_aqi', 'atmospheric particulate matter (PM2.5)', 'Atmospheric Particulate Matter, also known as atmospheric aerosol particles,\nare complex mixtures of small solid and liquid matter that get into the air.\nIf inhaled they can cause serious heart and lungs problem.']]

for option in column_description:

column_1 = f'{option[0]}_value'

column_2 = f'{option[0]}_category'

title_column = option[1]

this_df = (countries_cities[[column_1,column_2]].sort_values(by=column_1,ascending=False).head(25))

(this_df.reset_index().sort_values(by=column_1,ascending=True)

.plot(kind='barh',x='city',y=column_1,color=colour,legend=False))

this_median = countries_cities[column_1].median()

plt.axvline(x=this_median,ls=':',lw=2.5,color='white')

plt.title(f'25 areas with maximum values\nof {title_column}\n', fontsize=25, loc='left')

plt.xlabel(f'{title_column} value')

plt.ylabel("")

plt.show()

print(f'{option[2]}\n')

print(f'White dotted line: {this_median} - median value for {title_column} per city/area\nwithin the dataset.')

print('This means, 50% of cities have no more than this.')

print('\n')

display(this_df)

```

print("\n\nLet's summarise.")
print(f'Countries whose cities are listed in the top 25 of the highest {title_column}
level:')
for item in this_df.reset_index()['country'].unique():
    print(f'- {item}')
print("\nThe most polluted areas, five examples:")
for item in this_df.reset_index()['city'].to_list()[:5]:
    print(f'- {item}')
print("\nCategories for the selected 25 locations:")
for item in this_df.reset_index()[column_2].unique():
    print(f'- {item}')
print("\n")
print(100*_,' \n\n')
sns.boxplot(df['AQI Value'], color='Green')
plt.title('Air Quality Index', fontsize=50, color='Blue')
plt.xlabel('Air Quality', fontsize=30, color='Magenta')
plt.grid()
plt.show()
def remove_outlier(col):
    q1, q3 = col.quantile([0.25, 0.75])
    iqr = q3 - q1
    lower_range = q1 - 1.5 * iqr
    upper_range = q3 + 1.5 * iqr
    return lower_range, upper_range
lowage,uppage=remove_outlier(df["AQI Value"])
df["AQI Value"]=np.where(df["AQI Value"]>uppage,uppage,df["AQI Value"])
df["AQI Value"]=np.where(df["AQI Value"]<lowage,lowage,df["AQI Value"])
sns.boxplot(df['AQI Value'], color='Green')
plt.title('Air Quality Index', fontsize=50, color='Blue')
plt.xlabel('Air Quality', fontsize=30, color='Magenta')
plt.grid()
plt.show()
df['AQI Category'].value_counts().plot(kind='bar', color='Sandybrown')

```

```
plt.title('Air Quality Index',fontsize=50,color='Blue')
plt.xlabel('AQI Category',fontsize=30,color='Magenta')
plt.ylabel('AQI Value', fontsize=30, color='magenta')
plt.show()

sns.regplot(x='AQI Value', y='Ozone AQI Value', data=df)
plt.title('Regression Plot', fontsize=50, color='Blue')
plt.xlabel('Air Quality', fontsize=30, color='Magenta')
plt.ylabel('Ozone Air Quality', fontsize=30, color='Magenta')
plt.show()
```

B. SCREENSHOTS

	dtype	specimen	nunique	null_share
Country	object	Russian Federation	175	1.82%
City	object	Praskoveya	23462	0.00%
AQI Value	int64	51	347	0.00%
AQI Category	object	Moderate	6	0.00%
CO AQI Value	int64	1	34	0.00%
CO AQI Category	object	Good	3	0.00%
Ozone AQI Value	int64	36	213	0.00%
Ozone AQI Category	object	Good	5	0.00%
NO2 AQI Value	int64	0	59	0.00%
NO2 AQI Category	object	Good	2	0.00%
PM2.5 AQI Value	int64	51	383	0.00%
PM2.5 AQI Category	object	Moderate	6	0.00%

Data Type and their Unique values

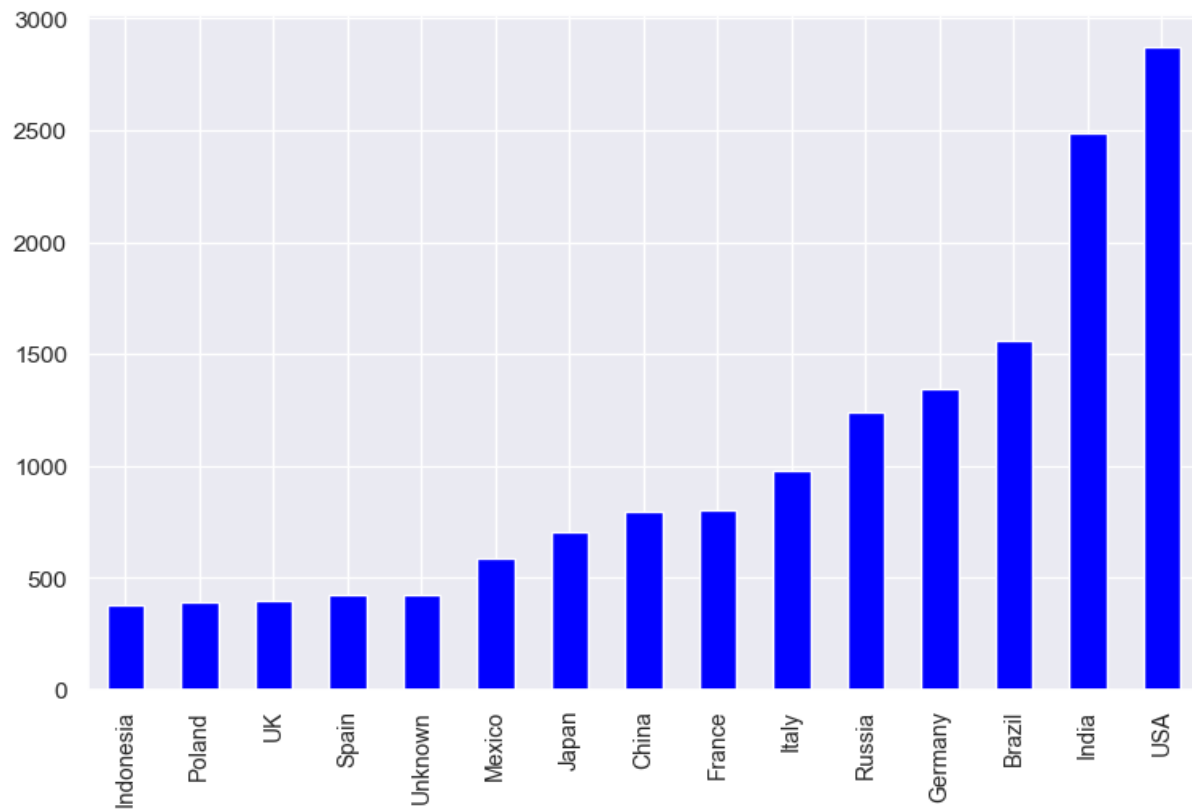
===== First Five Rows =====

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Russian Federation	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
1	Brazil	Presidente Dutra	41	Good	1	Good	5	Good	1	Good	41	Good
2	Italy	Priolo Gargallo	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
3	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
4	France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good

===== Last Five Rows =====

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
23458	India	Gursahaiganj	184	Unhealthy	3	Good	154	Unhealthy	2	Good	184	Unhealthy
23459	France	Sceaux	50	Good	1	Good	20	Good	5	Good	50	Good
23460	India	Mormugao	50	Good	1	Good	22	Good	1	Good	50	Good
23461	United States of America	Westerville	71	Moderate	1	Good	44	Good	2	Good	71	Moderate
23462	Malaysia	Marang	70	Moderate	1	Good	38	Good	0	Good	70	Moderate

Details on First and Last Rows



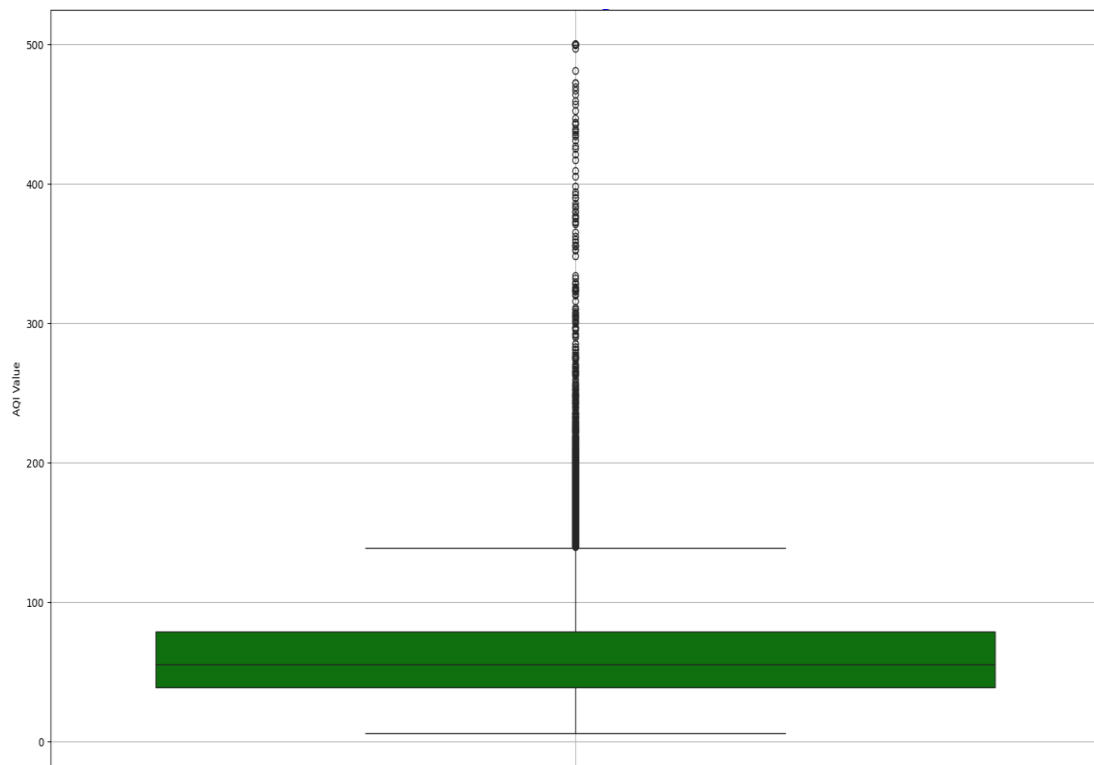
Well Represented Countries

Spain	87.53%	12.47%	0.00%	0.00%	0.00%	0.00%
Russia	82.59%	16.36%	0.81%	0.08%	0.08%	0.08%
UK	79.75%	20.25%	0.00%	0.00%	0.00%	0.00%
Brazil	72.02%	22.28%	2.50%	2.82%	0.38%	0.00%
Japan	61.54%	36.18%	2.28%	0.00%	0.00%	0.00%
Poland	59.38%	37.79%	2.83%	0.00%	0.00%	0.00%
Unknown	58.08%	33.49%	5.39%	2.81%	0.23%	0.00%
Germany	53.31%	46.62%	0.07%	0.00%	0.00%	0.00%
France	51.62%	43.89%	3.62%	0.87%	0.00%	0.00%
USA	34.85%	59.71%	4.77%	0.63%	0.00%	0.03%
Italy	28.91%	67.31%	3.58%	0.20%	0.00%	0.00%
Mexico	23.81%	47.45%	9.86%	14.63%	3.23%	1.02%
Indonesia	20.05%	47.23%	9.23%	19.26%	4.22%	0.00%
China	7.17%	26.79%	23.02%	36.98%	5.66%	0.38%
India	5.23%	23.51%	15.39%	44.25%	5.27%	6.35%
	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous

Heat Map based on AQI Values

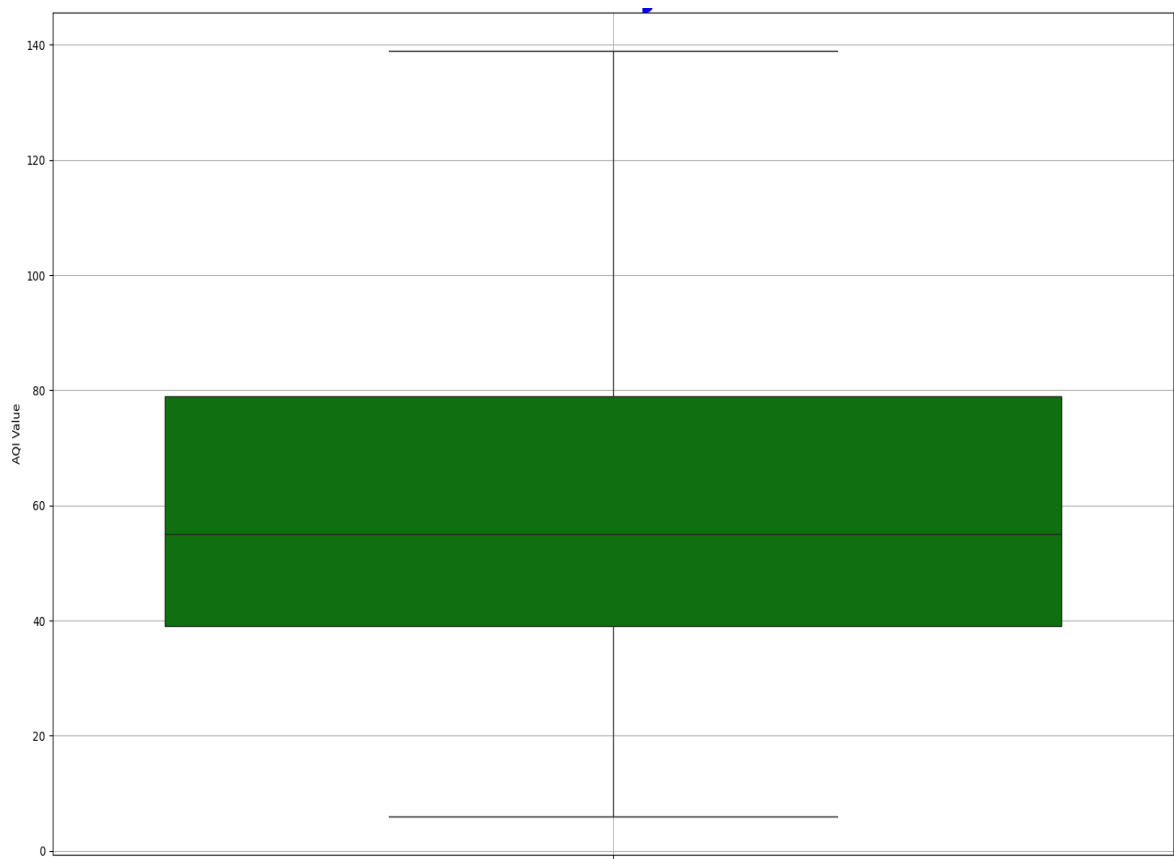
Spain	88.94%	11.06%	0.00%	0.00%	0.00%	0.00%
Russia	82.76%	16.20%	0.81%	0.08%	0.08%	0.08%
UK	79.75%	20.25%	0.00%	0.00%	0.00%	0.00%
Brazil	72.02%	22.28%	2.50%	2.82%	0.38%	0.00%
Japan	70.23%	27.78%	1.99%	0.00%	0.00%	0.00%
Unknown	60.42%	31.62%	5.39%	2.58%	0.00%	0.00%
Poland	59.38%	37.79%	2.83%	0.00%	0.00%	0.00%
Germany	53.31%	46.62%	0.07%	0.00%	0.00%	0.00%
France	52.37%	43.14%	3.62%	0.87%	0.00%	0.00%
Italy	38.61%	57.61%	3.58%	0.20%	0.00%	0.00%
USA	36.63%	57.97%	4.74%	0.63%	0.00%	0.03%
Mexico	23.81%	47.45%	9.86%	14.63%	3.23%	1.02%
Indonesia	20.05%	48.02%	9.76%	19.79%	2.37%	0.00%
China	7.92%	36.98%	25.91%	27.55%	1.38%	0.25%
India	5.23%	23.75%	15.72%	43.69%	5.95%	5.67%
	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous

Heat Map based on Atmospheric Particulate Matter



Air Quality

Box Plot with Outliers



Air Quality

Box Plot without Outliers

C. RESEARCH PAPER

Global Air Pollution Prediction using Decision Tree

Ashish Kumar

Department of Computer Science and
Engineering
Sathyabama Institute of Science and
Technology
Chennai, India
ashishkumar17012002@gmail.com

Jeslin Shanthamalar J

Department of Computer Science and
Engineering
Sathyabama Institute of Science and
Technology
Chennai, India
jeslin.cse@sathyabama.ac.in

Nancy Noella RS

Department of Computer Science and
Engineering
Sathyabama Institute of Science and
Technology
Chennai, India
nancynoella.cse@sathyabama.ac.in

Abstract—Air pollution is a worldwide environmental concern that significantly impacts the health and well-being of individuals across the globe. The impact of air pollutants on the health of human beings, ecosystems, and the climate has spurred significant concern and necessitated comprehensive analysis and understanding. This study provides an in-depth examination of global air pollution trends, sources, and their consequences, drawing from a vast and diverse dataset. The "Global Air Pollution Dataset" serves as the foundation for this analysis, offering a wealth of information that encompasses air quality measurements, emissions data, meteorological variables, and more. The dataset combines historical records with real-time updates, providing a rich resource for researchers, policymakers, and environmental enthusiasts. The proposed model achieved 98% in terms of accuracy using Decision Tree Algorithm to analyze the Pollutants as a public health concern that include particulate matter, carbon monoxide, nitrogen dioxide and ozone.

Keywords—Machine Learning, Deep Learning, Decision Tree Algorithm, Random Forest

I. INTRODUCTION

Air pollution is a vast global environmental challenge with the consequences for public health, ecosystems, and climate change. The presence of pollutants such as particulate matter, nitrogen oxides, carbon monoxide, and organic compounds in the atmosphere has led to a host of adverse effects, including respiratory diseases, reduced air quality, and altered weather patterns.

The issue of air pollution has emerged as one of the most environmental challenges confronting our planet. The degradation of air quality due to pollutants emitted from various human activities has consequences for both human health and the environment. From densely populated urban centers to remote rural areas, no corner of the globe remains untouched by the pervasive impacts. Due to this pollution now a days in clear dark sky at night we can't see our milky way galaxy, while in our history looking back our eyes can see the milky way galaxy without any equipments it didn't ends here animals and insects like beetles used to travel in the night sky straight without any hesitation. They were following a galaxy white strip-like structure. These days we can't be able to see because of pollution and it causes some myopia disease which affects human's eyes as well as animals.

The air pollution is multifaceted and complex, encompassing a spectrum of health effects, environmental degradation, and socioeconomic implications. Exposure to polluted air is linked or we can say that lead to health

problems, ranging from respiratory diseases and lungs cancer to neurological damages..

Air pollution exacts a toll on ecosystems, disrupting ecological processes and biodiversity. Acid rain, a consequence of air pollution caused by sulfur dioxide and nitrogen oxides emissions, poses a threat to forests, freshwater bodies, and aquatic life. Moreover, airborne pollutants can accumulate in soil and water, entering food chains and posing risks to wildlife and human populations.

The global challenge of air pollution requires concerted efforts at local, national, and international levels, encompassing policy interventions, technological innovations, and behavioral changes. Strategies to mitigate air pollution encompass a range of measures, including the adoption of cleaner energy sources, improved vehicle emissions standards, enhanced industrial pollution control, and sustainable land management practices.

II. RELATED WORK

Doreswamy et al. investigated the ML predictive models for forecasting Particulate Matter that have to be pollutants in the air. But according to the analysis they only went to taiwan for their analysis. Madan T et al. Air pollution is very dangerous and should be decreasing as fast as they could. These are the things they have to keep in their mind for prediction or for analysis. Madhuri VM et al. The presence of air pollutants in air is accessed. These can also be able to predict by using probability and statistics. So they used some of these algorithms and worked on that. C. Wu et al. To improve the air quality, He was trying to show that air pollution has gone through several stages in the evolution of single pollutants into composite pollutants which had been caused by the existence of many or several pollutants. Bhalgat P et al. Work as Monitoring and ensuring that good air quality has become a top priority for governments in urban areas. global warming, acid rains, or an increased number of patients.

G. K. Kang et al. Observing and maintaining air quality has become important in many places. The quality of air has been affected because of pollution. Y. Chen et al. Inhaling polluted air for a severe long time can lead to many physical and mental problems. K. Kujaroentavon et al. The goal for doing work is to keep rules on the basis of the quality of air. The results of this study were correctly classified into a training. W. Bin In The Air Quality Index of capitals, their cities to find the distribution across all cities. They used to know about the data by using some of the graphs and

analyzing them. K. Veljanovska and A. Dimoski They found that the pollution Affecting human's respiratory systems is a vast cause for increase in death rate and shortened lifespan and increased the risk of spreading diseases.

G. Parthasarathy and B. N. Chatterji It involves determining the density of observed data points with k as the number of the neighbors. K. H. Brodersen et al. It is used in the performance on information retrieval or checks the overall performance. On the basis of accuracy on two limitations. F. A. Gers as they have been trained in their neural network to predict a clean speech if there is some noise available in their background so they used a noise feature. The first method is Phase error based filtering which performs on the basis of time difference. The second method is Correlation shaping. M. Zhao and X. Li as this is based on the air quality but they have distributed it in a way of a smart city. By making a weather list system it will improve overall understanding of the environment.

V. M. Niharika and P. S. Rao Present in the atmosphere in a way that can harm humans, plants or animals and the environment will be damaged by those substances and it will show the effect on the weather or on the sudden climatic change. D. Mishra as they used the Neuro-Fuzzy model at that time so for the air quality forecasting they chose to deal with the Nitrogen dioxide (NO₂) this is the pollutant that had been chosen for the analysis. T. Lan et al. as they decided to work on the disease which is called as spread they used to store the data in the format of ionograms so they decided to train and process. Xiaosong Zhao et al. as they used deep learning to go for the prediction and then for classification, they decided to use a Support machine, Random forest and the RNN model they used in three different industrial places.

Venkat-Rao Pasupuleti et al. as it used to analyze the air quality so they used air quality Index for only India to and used an useful approach for resolving the quality analysis of air. Soubhik Mahanta et al. decided to do this prediction because they don't want any of the people to face danger or suffer from the pollution. Zhili Zhao et al. They decided to check for every hour, which means they will show the hourly prediction which means they have a huge impact on weather or pollution. Kostandina Veljanovska and Angel Dimovski as I understand in his paperwork he is using AI tools so that it can make decisions on its own not by a programmer that will feed all the data and say it to show the values. Kumar et al. The findings of the research offer practical insights for enhancing air quality monitoring and management strategies, showcasing their relevance in real-world scenarios.

III. DESCRIPTION

The Title of the dataset that is used in the proposed work is Global Air Pollution. The proposed work uses a dataset which contains 23463 instances with 12 attributes. These Columns are Country, City, AQI (Air Quality Index) Value, AQI Category, Carbon Monoxide gas, Ozone, Nitrogen Dioxide gas and Particulate Matter. These are so harmful if you make contact between them.

Describing proposed work of the program as histogram, Bar Graph, Regression Plot, Box plot, Heat Map is used. By talking about Categories it depends on AQI conditions like AQI is good condition, Moderate condition, Unhealthy for the sensitive groups of situation, Very Unhealthy, or Unhealthy, or Hazardous disease or situation. By using a heat map, we can verify the level of air pollution is present in that

place or the country. It will help in predicting the air pollution according to the data provided. Also by using Hierarchical Clustering to show the air quality means air pollution level according to country based.

Actually by using median and mode values of AQI taken for prediction here. Also by using Heat map here we separated which gasses are polluting which of the specific places we like to take first from the proposed system by taking Ozone to see which places are polluted in what ways like good, moderate, Unhealthy or Hazardous.

IV. IMPLEMENTATION

This paper presents a technique for global air pollution. Structural changes in the architecture of global air pollution prediction. As part of data collection this data is collected based on countries and cities and some of the pollutant gasses and their AQI Values with their Category. The name of the selected gasses which proposed work taken is:

- i) Carbon Monoxide gas (CO)
- ii) Ozone layer (O₃)
- iii) Nitrogen Dioxide gas (NO₂)
- iv) Particulate Matter (PM_{2.5})

Before Starting every city has their own AQI values and their Categories but here to say that the proposed work has to see how it will affect the pollution level in the country. AQI Values of every gasses plays a crucial role in the prediction.

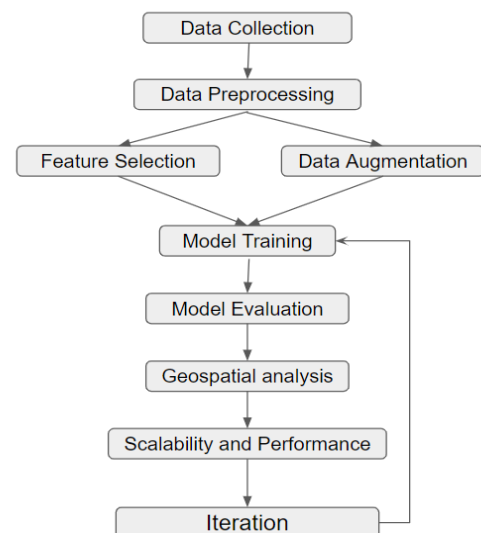


Fig. 1. Architecture Diagram

Carbon Monoxide [CO]

Carbon monoxide (CO) is a gas that lacks color and odor. It is emitted by sources like kerosene and gas space heaters, along with gas stoves, which can degrade indoor air quality. Exposure to CO can lead to symptoms such as dizziness, loss of consciousness, and even death.

Ozone [O₃]

Ground-level ozone, produced through chemical reactions involving nitrogen oxides and specific organic compounds, presents health hazards and serves detrimental purposes. It can diminish lung function and exacerbate conditions like asthma. It damages vegetation during the growing season.

Nitrogen Dioxide [NO₂]

Nitrogen Dioxide (NO₂) is classified as one of the nitrogen oxides. While it can naturally enter the air from the

stratosphere, at ground level, NO₂ primarily originates from emissions of cars, trucks, buses, and power plants.

Particulate Matter [PM2.5]

Atmospheric Particulate Matter consists of intricate combinations of small solid particles as well as liquid particles that are in the air. Inhalation of these particles can lead to severe heart and lung issues.

A. Data Processing

First we will divide all the numeric values and non-numeric values. In numeric values these aqi_value, co_aqi_value as a numeric value, ozone_aqi_value, no2_aqi_value there where ever is value those belong to numerical categories and pm2_5_aqi_values. For non-numeric values are aqi_category, co_aqi_category these are the one who doesn't contains numeric, ozone_aqi_category, no2_aqi_category these only contains object and pm2_5_aqi_category now here the density of values in each numeric column is here. So, these are the densities in each of the numeric columns.

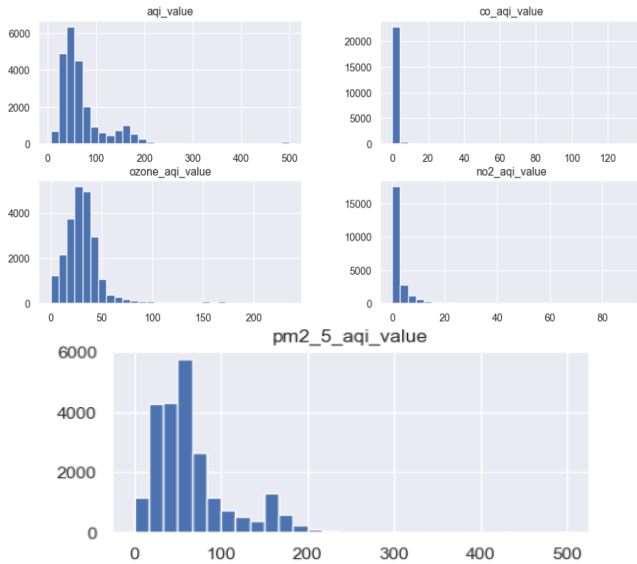


Fig. 2. Density of Polluted Air

B. Feature Selection and Data Augmentation

count	176.000000
mean	133.312500
std	355.508745
min	1.000000
25%	10.750000
50%	28.500000
75%	81.000000
max	2872.000000

Fig. 3. Country details

For the feature selection proposed work have taken the country column as in Fig.3, so now by describe the country we get the data about the country

Now, for the proposed work going to check the values present in the or we can say the name of the country which is greater or larger we have to reduce the size of the name of the country as in short form. Here the proposed work going to show the top or most well represented country that are present in the data is

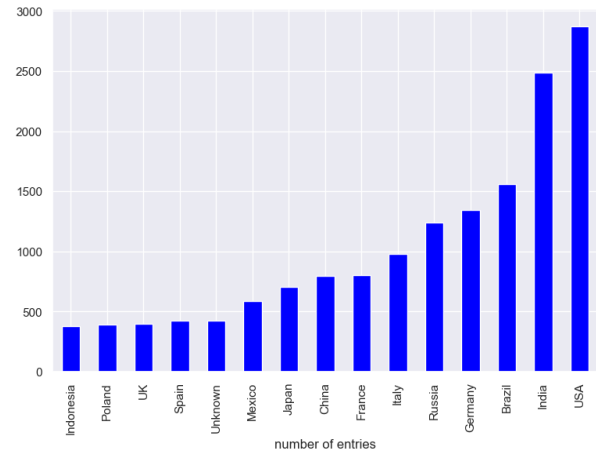


Fig. 4. Well Represented Country

Proposed system had 23462 locations in 175 countries. There is too much information to fit onto the screen without dividing it into smaller pieces, focusing on something more particular or doing something else about it. To do something about it, first, work took a more detailed look at the most represented countries within the dataset. It shows the well represented details of the country regarding different air pollution measurements. In order to do so, the proposed work used heat maps to show the air's quality index of countries.

Spain	87.53%	12.47%	0.00%	0.00%	0.00%	0.00%
Russia	82.59%	16.36%	0.81%	0.08%	0.08%	0.08%
UK	79.75%	20.25%	0.00%	0.00%	0.00%	0.00%
Brazil	72.02%	22.28%	2.50%	2.82%	0.38%	0.00%
Japan	61.54%	36.18%	2.28%	0.00%	0.00%	0.00%
Poland	59.38%	37.79%	2.83%	0.00%	0.00%	0.00%
Unknown	58.08%	33.49%	5.39%	2.81%	0.23%	0.00%
Germany	53.31%	46.62%	0.07%	0.00%	0.00%	0.00%
France	51.62%	43.89%	3.62%	0.87%	0.00%	0.00%
USA	34.85%	59.71%	4.77%	0.63%	0.00%	0.03%
Italy	28.91%	67.31%	3.58%	0.20%	0.00%	0.00%
Mexico	23.81%	47.45%	9.86%	14.63%	3.23%	1.02%
Indonesia	20.05%	47.23%	9.23%	19.26%	4.22%	0.00%
China	7.17%	26.79%	23.02%	36.98%	5.66%	0.38%
India	5.23%	23.51%	15.39%	44.25%	5.27%	6.35%
	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous

Fig. 5. Air quality index of most risky countries

Here in Fig.5 in this table everyone can see that there is Quality of air where and in which country how much pollutants are there. This one here only shows the Air Quality index of the country without based on the different gasses. Like this there are more heat maps proposed work have created and those all are based on the different gasses and countries. Each specific gas has their own heat maps based on the pollution level. We can clearly see that the condition

is good, Moderate, Unhealthy, Hazardous all the data have been distributed among the others.

Here we can easily evaluate the data which is present in the dataset like here in the proposed work used boxplot to show the air quality index by using AQI Value is

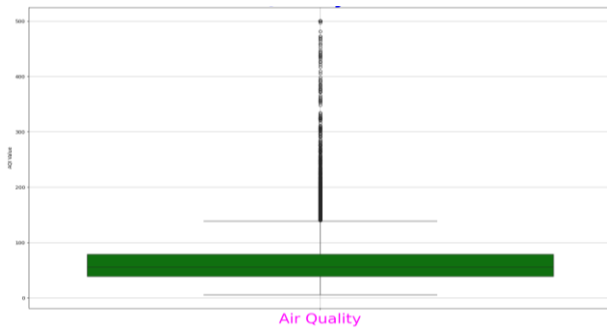


Fig. 6. Box plot for air quality index with outliers

Here we saw that there are many outliers in the box plot so here we can't be able to get the accurate measurement or clear distribution of the Quality of air. So by using Interquartile range proposed work fixed the issue. Here proposed work used the formula for removing the outliers which is as follows:

First Quartile Q1 (25th percentile) = 39.0

Median Q2 (50th percentile) = 55.0

Third Quartile Q3 (75th percentile) = 79.0

$IQR = Q3 - Q1 = qn(0.75) - qn(0.25)$

$79.0 - 39.0 = 40.0$

The upper quartile (Q3) spans a distance equal to 1.5 times the interquartile range (IQR) from the third quartile. This measurement extends to the highest recorded data point within this interval from the dataset. so,

$< Q1 - 1.5 * IQR$

$39.0 - 1.5 * 40.0 = -21.0$

A distance equal to 1.5 times the interquartile range (IQR) is subtracted from the lower quartile (Q1). This distance is then extended downward to the lowest observed data point within this range from the dataset.

$> Q3 + 1.5 * IQR$

$79.0 + 1.5 * 40.0 = 139$

So, after using that formula for the boxplot we get the clarity for the Air Quality index value.

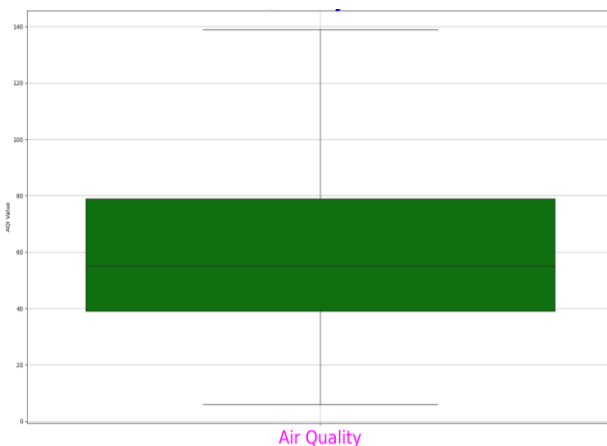


Fig. 7. Box plot for air quality index without outliers

C. Model Training

Here in this Global air pollution proposed work there used the dataset to train models of air pollution. By using sklearn.model_selection proposed work used the dataset to train and test the model by splitting it into several parts. At first it took 20% percent of the data to train and then test like we can say that here in the proposed work used algorithms to evaluate the data like Decision tree algorithm.

Prediction	Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Actual						
Good	8166	0	0	0	0	0
Hazardous	0	138	0	0	0	0
Moderate	0	0	7260	0	0	0
Unhealthy	0	0	0	1703	0	0
Unhealthy for Sensitive Groups	0	0	0	0	1299	0
Very Unhealthy	0	0	0	0	0	204

Fig. 8. Training data results

Prediction	Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Actual						
Good	2042	0	0	0	0	0
Hazardous	0	34	0	0	0	0
Moderate	0	0	1815	0	0	0
Unhealthy	0	0	0	426	0	0
Unhealthy for Sensitive Groups	0	0	0	0	325	0
Very Unhealthy	0	0	0	0	0	51

Fig. 9. Testing data results

Decision Rule: if $X_i \leq T$, go to left child; otherwise, go to right child

Entropy: $Entropy(S) = -n \sum_{i=1} P(x_i) \log_2 P(x_i)$

Gain(S,Attribute) = Entropy(S) - $\sum |S_v|/|S|$ (Entropy(S_v))

where,

$v = \{\text{attribute values}\}$

On the Gain value we can get the arrangement of a decision tree algorithm.

By splitting I got the values as

{ 'min_samples_split': 27, 'min_samples_leaf': 46, 'max_leaf_nodes': 40, 'max_depth': 10, 'criterion': 'entropy' }

By using these splitting over decision tree proposed work got the value as for testing data of these 20% training the data proposed work got,

Prediction	Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Actual						
Good	2039	0	3	0	0	0
Hazardous	0	19	0	15	0	0
Moderate	15	0	1793	0	7	0
Unhealthy	0	0	0	421	5	0
Unhealthy for Sensitive Groups	0	0	0	9	315	1
Very Unhealthy	0	2	0	9	0	40

Fig. 10. Testing data for accuracy

The accuracy of the Decision tree algorithm is approximately 98%. After using Randomized algorithm and Decision Tree

Algorithm proposed work found that Decision Tree Algorithm is the best algorithm for the dataset.

D. Model Evaluation

In the proposed work that explored all the gasses and on the basis of the good, moderate, unhealthy, very unhealthy, hazardous these are the conditions of the gasses all over the places globally.

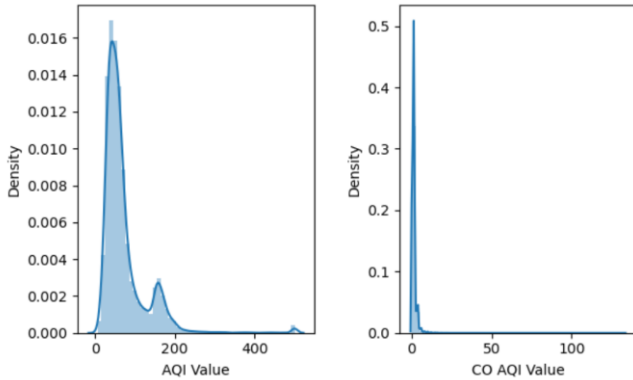


Fig. 11. Values based on polluted gasses

In the first view we can see that AQI Value is given based on a country and city. The order to analyze the graph is 'Moderate', 'Good', 'Very Unhealthy', 'Unhealthy for Sensitive Groups', 'Unhealthy', 'Hazardous'.

In the second one we can clearly analyze that this AQI value is dependent on the Carbon Monoxide gas in the country and its order to see the gasses is 'Good', 'Unhealthy for Sensitive Groups', 'Moderate'.

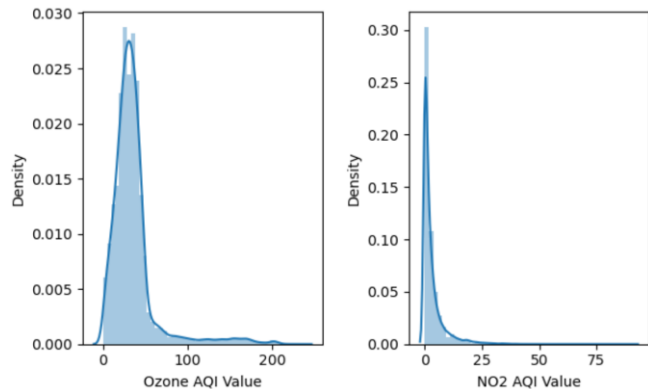


Fig. 12. Values based on the Ozone and Nitrogen Dioxide gasses

In Fig.12 we can clearly analyze that this AQI value is depend on the Ozone gas in the country and its order to see the gasses is 'Good', 'Moderate', these are normal 'Unhealthy for Sensitive Groups', 'Unhealthy', these can go to severe 'Very Unhealthy', if unhealthy is there then big or serious issues.

In Fig.12 We can clearly analyze that this AQI value is dependent on the Nitrogen Dioxide gas in the country and its order to see the gasses is 'Good', 'Moderate'. There is no large amount of Pollution caused by NO2.

In Fig.13 we can clearly analyze that this AQI value is depend on the Particulate Matter 2.5 gas in the country and its order to see the gasses is 'Moderate', 'Good', 'Unhealthy for Sensitive Groups', 'Very Unhealthy', 'Unhealthy', 'Hazardous'.

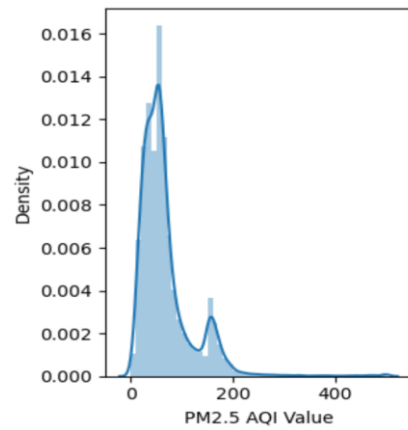


Fig. 13. Values based on Particulate Matter

TABLE I. COMPARISON BETWEEN TWO ALGORITHMS

Algorithms	Accuracy
Random Forest	82.35%
Decision Tree	98.59%

So, accuracy of the Decision Tree Algorithm is more than the Random Forest Algorithm so here the proposed work preferred to use the Decision Tree Algorithm for getting the best suitable Accuracy.

V. COMPARATIVE ANALYSIS

Here the performance measures of existing works are given below:

TABLE II. COMPARATIVE ANALYSIS OF EXISTING METHOD VS PROPOSED METHOD

SI	Existing Method	Algorithm	Accuracy
1	K. Hu (2017)	SVR	64.2%
2	Madhuri VM (2020)	SVM	75%
3	Venkat Rao Pasupuleti (2020)	Random Forest	79%
4	Soubhik Mahanta (2019)	Extra Trees	85.3%
5	Y. Wang (2019)	BP Neural Network	91%
6	Kumar. (2023)	KNN	89%
7	Proposed Work	Decision Tree	98%

The proposed system got 98% accuracy by using Decision tree Algorithm which is quite higher than existing work.

Pollution Estimation From Fixed and Mobile Sensors K. Hu et al. The effectiveness of the system may be constrained to Sydney or comparable urban settings, potentially diminishing its performance when applied to diverse geographical areas. Madhuri VM, Samyama GGH complex algorithms like Random Forest may be prone to overfitting, meaning they could excessively fit the training data and perform poorly

Venkat Rao Pasupuleti et al. Utilizing sensors on the Arduino Uno platform might restrict the monitoring system's coverage area, potentially resulting in incomplete data collection over a broader geographical expanse. This limitation could compromise the system's ability to capture a comprehensive picture of air quality across a large area. Soubhik Mahanta et al. Overfitting arises when a model becomes too closely attuned to the intricacies of the training data, resulting in diminished performance when applied to new or unseen data. Y. Wang, T. Kong. The comparison of the improved algorithm to other classification prediction methods might lack comprehensiveness

VI. CONCLUSION

In conclusion, predicting global air pollution is a crucial endeavor that requires ongoing efforts and advancements in technology, data analysis, and international collaboration. As our world continues to industrialize and urbanize, the challenges associated with air quality are likely to take proactive measures to manage the impacts on human health, ecosystems, and climate. The proposed method predicted the air pollution with 98% accuracy using a decision tree algorithm.

Public awareness and engagement are essential components of any successful air pollution prediction and control strategy. Educating communities as to the health risks associated with the poor air quality. In conclusion, the ongoing commitment to research, technological innovation, global collaboration, and public involvement is vital for developing robust predictive models and effective strategies to decrease the impacts of air pollution on a global scale.

Public awareness and engagement are essential components of any successful air pollution prediction and control strategy. Educating communities about health risks as with a poor quality of air, that encourage individuals to contribute to a cleaner and healthier environment. ongoing commitment to research, technological innovation, global collaboration, and public involvement is vital for developing robust predictive models and effective strategies to address and decrease the impacts of air pollution on a global scale.

REFERENCES

- [1] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning which has been based on regression models. *Procedia Comput Sci* 171:2057–2066.
- [2] Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms. In: 2nd international conference on advances in computing, communication control and networking (ICACCT) pp 140–145
- [3] K. Hu, A. Rahman, H. Bhugubanda and V. Sivaraman, "HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors," in *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3517–3525, 1 June 1, 2017, doi: 10.1109/JSEN.2017.2690975.
- [4] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning and supervised learning approach. *Int J Sci Technol Res* 9(4):118–123.
- [5] C. Wu, W. Hu, M. Zhou, S. Li, and Y. Jia, "Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China," *Atmos. Environ.*, vol. 203, pp. 172–182, Apr. 2019.
- [6] Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. *Int J Comput Appl Technol Res* 8(9):367–370.
- [7] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *Int. J. Environ. Sci. Develop.*, vol. 9, no. 1, pp. 8–16, Jan. 2018.
- [8] Y. Chen, L. Wang, F. Li, B. Du, K.-K. R. Choo, H. Hassan, and W. Qin, "Air quality data clustering using EPLS method," *Inf. Fusion*, vol. 36, pp. 225–232, Dec. 2016.
- [9] K. Kujaroentavon, S. Kiattisin, A. Leelasantham, and S. Thammaboosadee, "Air quality classification in Thailand based on decision tree," in *Proc. 7th Biomed. Eng. Int. Conf.*, Fukuoka, Japan, Nov. 2014, pp. 26–28.
- [10] W. Bin, "Analyze the spatial-temporal characteristics of air pollution in China by using air pollution index (API)," M.S. thesis, Ocean Univ. of China, Qingdao, China, 2008
- [11] Ramani RG, Shanthamalar JJ. Automated image quality appraisal through partial least squares discriminant analysis. *Int J Comput Assist Radiol Surg.* 2022 Jul;17(7):1367–1377.
- [12] K. Veljanovska and A. Dimoski, "Air quality index prediction using simple machine learning algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 7, no. 1, pp. 25–30, Jan./Feb. 2018.
- [13] G. Parthasarathy and B. N. Chatterji, "A class of new KNN methods for low sample problems," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 3, pp. 715–718, May 1990 doi: 10.1109/21.57285.
- [14] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The binormal assumption on precision-recall curves," in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 4263–4266
- [15] F. A. Gers, "Long short-term memory in recurrent neural networks," M.S. thesis, Dept. Comput. Sci., Swiss Federal Inst. Technol., Lausanne, EPFL, Switzerland, 2001.
- [16] M. Zhao and X. Li, "An application of spatial decision tree for classification of air pollution index," in *Proc. 19th Int. Conf. Geoinform.*, Shanghai, China, Jun. 2011, pp. 1–6.
- [17] V. M. Niharika and P. S. Rao, "A survey on air quality forecasting techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 103–107, Jan. 2014.
- [18] D. Mishra, "Neuro-fuzzy approach to forecast NO2 pollutants addressed to air quality dispersion model over Delhi, India," *Aerosol Air Qual.*, vol. 16, no. 1, pp. 166–174, Jan. 2016.
- [19] T. Lan, Y. Zhang, C. Jiang, G. Yang, and Z. Zhao, "Automatic identification of spread F using decision trees," *J. Atmos. Sol.-Terr. Phys.*, vol. 179, pp. 389–395, Nov. 2018.
- [20] Xiaosong Zhao, Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan Ze University, A Deep Recurrent Neural Network for Air Quality Classification, 2018.
- [21] Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth and Hari Kiran Reddy, Air Quality Prediction Of Data Log By Machine Learning, 2020, IEEE
- [22] Soubhik Mahanta, T. Ramakrishnudu, Rajat Raj Jha and Niraj Tailor, Urban Air Quality Prediction Using Regression Analysis, 2019, IEEE
- [23] Zhili Zhao, Jian Qin, Zhaoshuang He, Huan Li, Yi Yang and Ruisheng Zhang, Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China, Environmental Science and Pollution Research, 2020
- [24] Kostandina Veljanovska1 & Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms, 2018, International Journal of Emerging Trends & Technology in Computer Science (IJETCS).
- [25] Kumar, K., Pande, B.P. Air pollution prediction with machine learning: a case study of Indian cities. *Int. J. Environ. Sci. Technol.* 20, 5333–5348 (2023).
- [26] Y. Wang and T. Kong, "Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid," in *IEEE Access*, vol. 7, pp. 172892–172901, 2019,

PAPER ACCEPTANCE

Acceptance letter for NMIC24051 Inbox x



NMIC 2024 <nmic.2024.02@gmail.com>
to me ▾

Sun, Feb 25, 12:50 PM ☆ ☺ ↶ ⋮

Acceptance letter for NMIC24051

Dear Authors,

"Congratulations"

Your paper of **Paper ID: NMIC24051** is accepted for NMIC-2024 publications in proceedings (journals with **minor revisions**). Please incorporate the reviewers' comments and re-submit the papers in the name of paper ID only e.g. NMIC24001 (docx file) with Plagiarism report (pdf file with same paper ID) by authenticated software (iThenticate, etc) within 3 days of this acceptance letter.

Registration Form, NEFT details & Copyright Form scan copy sent by e-mail within 2 days of receiving this acceptance letter and hard copy carry at the time of presentation. Authors can submit paper fees in the form of IMPS/NEFT/RTGS/ Phone Pay as per his/her convenience. All papers sent on conference e-mail: nmic.2024.02@gmail.com, admnm.isve.raschi@gmail.com, iete.isve.vlsi@gmail.com only. Conference presentations are **online mode** only. If you want special support in publication/accommodation, please register in a special academia/industry Gold Special or Diamond special domain. Letter for oral presentation will be sent after your successful registration.

Papers will be accepted in word format only. Previous conference published articles are available on given link: www.springer.com/gp/book/9789811029981 and <http://link.springer.com/article/10.1007/s00542-016-2873-8>.

The accepted papers will be published in Peer reviewed, UGC CARE, Scopus, Web of Science, SCI and some other indexed journals. (Name of Journals: 1-IETE Journal of Research (<http://mc.manuscriptcentral.com/ijr>), 2-IETE Technical Review, 3-JOLPE, 4-IETE Journal of Education, 5- International Journal of Microsystems and IoT, 6-Analog Integrated Circuits and Signal Processing, 7- IJCA, 8-TELKOMNIKA, 9- Microsystems Technologies, 10. IEEE- ACCESS, 11. Journal of Microprocessors and Microsystems, 12. Defense Science Journal, 13. Journal of Institutions of Engineers Series B, 14. Int. Journal of Communication Networks & Distributed Systems, Journal of Personal Communications, etc). **Please take care that plagiarism should not be more than 10-15%.**

ISVE Publications: NCCS-2015, MCCS-2015, NCCS-2016& MCCS-2017, NCCS-2017, MCCS-2018, NCCS-2018, MCCS-2019, NCCS-2019, MCCS-2020, MCMII-2020 accepted, presented and plagiarism free articles have been published in LNEE, Scopus book series of Springer and now all articles are available online on Springer website as well as ISVE website. As per request of several authors from India and abroad ISVE conference articles from NCCS-2020 are publishing in above said journals as per quality contents of articles with peer review support from experts. Extended version article published in listed Scopus & SCI journals. For more details you can visit the link: <https://www.isve.in/page.aspx?page=adminisve&pid=27>

ISVE Awards: Outstanding papers presenters will get the ISVE Best Paper Awards in each Technical Session recommended by review committee on the basis of novelty and will also get the opportunity for Cash Prize and good quality SCI Journal publications: <https://www.isve.in/page.aspx?page=adminisve&pid=27>

PAYMENT PROOF

	Transaction Successful 27 February 2024 at 10:13 PM
Transaction ID T2402272213535622465023 COPY	
Paid to	
	INDIAN SOCITY FOR VLSI E... UCO Bank - *****1831 ₹ 7,125
Debited from	
	XXXXXX3521 UTR: 405825432431 ₹ 7,125