Project Synopsis

on

**Pollution Tracking System**

Submitted as a part of course curriculum for

**Bachelor of Technology**

in

**Computer Science**



**Submitted by**

Nitesh Maurya

(1900290120073)

Meetika Jain

(1900290110060)

Sidhant Arya

(1900290120115)

**Under the Supervision of**

Prof. Akanksha

(Assistant Professor)

**KIET Group of Institutions, Ghaziabad**

**Department of Computer Science**

**Dr. A.P.J. Abdul Kalam Technical University**

**2022-2023**

**TABLE OF CONTENTS**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TITLE PAGE .................................................................................................................... | 1 | |  | | |  | |  | |
| DECLARATION .............................................................................................................. | 3 | |  | | |  | |  | |
| CERTIFICATE …........................................................................................................... | 4 | |  | | |  | |  | |
| ACKNOWLEDGEMENT.................................................................................................. | 5 | |  | | |  | |  | |
| ABSTRACT...................................................................................................................... | 6 | |  | | |  | |  | |
|  |  | |  | | |  | |  | |
| CHAPTER 1 INTRODUCTION | 7-9 | |  | | |  | |  | |
| 1.1         Introduction ……………………................................................... | 7 | |  | | |  | |  | |
| 1.2 Problem Statement.……………………..................................... | 8 | |  | | |  | |  | |
| 1.3         Objective………………………………………………………… | 9 | |  | | |  | |  | |
| CHAPTER 2 LITERATURE REVIEW…………………………………………….... | 10-11 | |  | | |  | |  | |
| CHAPTER 3 PROPOSED METHODOLOGY …………………………………........ | 12-13 | |  | | |  | |  | |
| 3.1 Flowchart | 12 | |  | | |  | |  | |
| 3.2 Algorithm Proposed | 13 | |  | | |  | |  | |
| CHAPTER 4 TECHNOLOGY USED ………..………………………..………………. | 14 | |  | | |  | |  | |
| CHAPTER 5 CONCLUSION …....................................................................................... 15 | |  | |  | | |  | |  | |
| REFERENCES…............................................................................................................ 16-17 | |  | | |  | |  | |  | |

|  |
| --- |
|  |

**DECLARATION**

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Name: Meetika Jain

1900290110060

Nitesh Maurya

1900290120073

Sidhant Arya

1900290120115

Date: 13/12/2022

**CERTIFICATE**

This is to certify that Project Report entitled “**Pollution Tracking System**” which is submitted by **Meetika Jain, Nitesh Maurya, Sidhant Arya** in partial fulfilment of the requirement for the award of degree B. Tech. in Department of Computer Science of Dr A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date: 13/12/2022 **Supervisor Signature**

Prof. Akanksha

(Assistant Professor)

**ACKNOWLEDGEMENT**

It gives us a great sense of pleasure to present the synopsis of the B. Tech Major Project undertaken during B.Tech. Fourth Year. We owe a special debt of gratitude to Prof.Akansha, Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his/her constant support and guidance throughout the course of our work. His/her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavours have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Dr. P. K Singh, Head of the Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Date: 30/08/2022

Names: Meetika Jain

1900290110060

Nitesh Maurya

1900290120073

Sidhant Arya

1900290120115

**ABSTRACT**

# In the process of making countries developed, the environment is often neglected. Various inventions made to help as well as provide comfort to mankind such as automobiles, air conditioners, etc.­­ led to the deterioration of the environment and the air quality. The major air pollutants are Carbon Monoxide (CO), Nitrogen Dioxide (NO2), particulate materials (PM2.5, SPM, and RSPM), Sulphur Dioxide (SO2), greenhouse gases, and Ozone(O3). These gases cause various respiratory diseases (such as asthma, and pulmonary cancer), Leukemia, Mesothelioma (a type of lung cancer), and premature deaths. In this paper, we have analyzed the datasets of previous years containing the values of various air pollutants such as NO2, SO2, SPM, RSPM and PM2.5 of years ranging from 1998 to 2020. The air quality indices (AQI) are then calculated using these values of air pollutants in the dataset and the calculated air quality indices (AQI) are included in our dataset to determine the air quality indices of future years in different cities of India. Supervised machine learning algorithms such as Linear Regression, Logistic Regression, Decision Tree and Random Forest are used for model training and determination of the air quality in the future years. The accuracy of different classification algorithms such as Logistic Regression, Decision Tree and Random Forest are also calculated to determine the most accurate algorithm for the prediction purpose.

**CHAPTER-1**

* 1. **INTRODUCTION**

The air gets polluted by harmful gases, solid and liquid substances suspended directly into the air. The biggest contributors to air pollution include industrial processes, the burning of firewood for home heating, the creation of electricity and heat, and transportation. The most harmful contaminant in the last years has been suspended materials. [14]. These substances can cause adverse effects on the environment and cause negative impacts on the health of humans. Exposure to air pollution causes short-term effects such as Bronchitis and Pneumonia and long-term effects such as heart diseases, lung cancer, and respiratory diseases like asthma on human health [15]. Pollution is responsible for approximately 9 million deaths per year, corresponding to one in six deaths worldwide [16]. Air pollution has badly affected not only the health of living beings but has also damaged the environment. Some of the harmful effects of air pollution on the environment are acid rain, eutrophication, haze, ozone depletion, crop and forest damage, and global climatic change [15]. The economic loss from premature deaths caused due to air pollution is quite high in India, equivalent to 1.36% of India's GDP in 2019. Health-care cost for the treatment of diseases caused due to air pollution is another source of economic loss. Based on National Health Accounts data, it is estimated that the total healthcare cost in India in 2019 to be $103·7 billion. Air pollution was responsible for 11.5% of the diseases in India in 2019, health-care costs for these diseases would be $11·9 billion (or 0·44% of India's GDP)[17].

It is important to closely investigate the patterns of air pollution in India and the states that are causing this severe problem. The aim is to analyze the air pollution trends of the past years and predict the rate of increase in pollution in the future years, which is brought out in this paper. We have studied the air pollution patterns from 1987 - 2022 within different states of India to find pollutants such as NO2, SO2, PM2.5, RSPM, and SPM which contaminate the air and cause various respiratory diseases. The detection methodologies of respiratory diseases like asthma are also discussed here.

* 1. **PROBLEM STATEMENT**

1. Air pollution has become major problem for every nation, whether it is developed country or developing country. Health issue have been growing rapidly especially in urban areas of developing countries where industrialization and growing number of vehicles leads to release of lot of air pollutants.
2. Adverse effects of pollution can cause allergic reactions such as irritation of the throat, eyes, and nose as well as some serious problems like bronchitis, heart diseases, pneumonia, lung, and aggravated asthma.
3. It is necessary to supervise air quality and keep it under control for a superior future and healthy living for all. With the Industrialization and with the increase in the vehicles on road the atmospheric conditions have considerably affected.
4. Testing of monitoring data allows us to assess how bad air pollution is from day to day. Due to resilience and high accuracy results of supervised machine learning algorithms, we are going to use these algorithms to make a model which will determine the extent to which air will get polluted within different cities in India.
5. The accuracy of different classification algorithms such as Logistic Regression, Decision Tree and Random Forest are also calculated to determine the most accurate algorithm for the prediction purpose.
   1. **OBJECTIVES**

Maintaining adequate air quality possesses a global challenge to governments and citizens. The deteriorating air quality has essentially caused governments globally to invest in multi-billion sums in policymaking and solution stratification to address the problem. Air pollution is caused by particulate matter emitted from industries, cars, machinery, waste recycling, industrial practices, and household. Some of the notable pollutants are the dust of heavy metals, carbon monoxide, ozone, carbon dioxide, nitrogen dioxide, suspended particulate matter, hydrogen fluoride oxides of sulphur, and others. These pollutants get into the atmosphere and cause severe health and environmental effects. In a survey conducted in 2016, the World Health Organization (WHO) stated that air pollution, specifically of the ambient (outdoor) origin, is the estimated cause of 4.2 million premature deaths annually worldwide. Air pollution impacts our lives and generations yet born; hence, we should be cautious and control harmful emissions of particulate matter into the atmosphere. A step in achieving this goal is by designing a model using machine learning algorithms with the following objectives,

* Analysis and pre-processing of the datasets of years ranging from 1998 to 2020 to train the model.
* Different supervised machine learning algorithms such as Linear Regression, Logistic Regression, Decision Tree and Random Forest are used for model training and determination of the air quality in future years in different cities of India.
* The accuracy of different supervised machine learning algorithms is calculated to determine if the results obtained by that algorithm are accurate or not.
* Comparing various classification machine learning algorithm such as Logistic Regression, Decision Tree and Random Forest to determine the most accurate algorithm for prediction purpose.

**CHAPTER-2**

**LITERATURE REVIEW**

According to V. Mijakovski and M. Lutovska [1], the main sources of air pollution include transportation, industrial operations, the production of power and heat, as well as the burning of wood for home heating. Suspended particles have historically been the most serious contaminant. With an estimated 12% of all deaths in 2019 being related to outdoor and indoor air pollution, it is clear that air pollution plays a significant role in the worldwide burden of disease Michael Brauer et al. [3]. Cardiovascular illnesses are thought to be the cause of 50% of the anticipated 6.7 million fatalities from air pollution in 2019, and air pollution is responsible for approximately 20% of heart disease deaths globally. Additionally, air pollution was listed as the fourth greatest risk factor. Madhuri VM et al. [2] used supervised learning algorithms, including LR, SVM, DT, and RF, for the prediction of air quality and found that RF performed better than Support Vector Machine, Decision Tree, and Linear Regression with the best accuracy. According to Ashima Tyagi et al. [4] research, India is one of the top two nations with the highest concentrations of PM2.5 and PM10 airborne particles. Additionally, the data reveals that Uttar Pradesh, Bihar, Rajasthan, and Haryana are the Indian states with the most toxic air. Delhi, however, is still not even among the top 10 cities with the worst particulate matter pollution. Kshitij Tripathi and Pooja Pathak [5] used deep learning techniques for air pollution, and this paper shows the area in which we are lacking (such as consideration of spatio-temporal information while predicting air pollution) in terms of predicting air pollution in India. This paper also shows that the widely adopted and well-liked approach for forecasting air pollution is CNN-LSTM. Bhishma Tyagi et al. [6] incorporated Menn-Kendall trend analysis was used to comprehend Aerosol Optical Depth trends over India and to determine variations in the air pollution pattern during the time of covid by percentage changes in AOD, Nitrogen dioxide, and Nitrogen dioxide. According to the findings, there are clusters of coal-fired power stations in eastern India with greater air pollution levels. Our findings indicate that, considering the prior anticipated studies, eastern India will experience higher levels of air pollution, making it a new hotspot location for air pollution with the highest magnitudes. Mriduchhanda Chattopadhyaya et al. [7] based on information from 557 rural Indian families, discovered that the percentage of days that households used dirty fuel had a negative and substantial relationship with subjective probabilistic estimates of getting sick because of using it. Girish Agrawal et al. [8] research demonstrates that assessing air quality in small cities can be done with inexpensive sensors, and it offers relatively cheap solutions for cities to significantly reduce the number of air pollution-related illnesses and deaths, as well as the negative environmental effects of cities on a per-capita basis. These sensors were used to track trends in particulate air pollution over a year in three small Indian cities: Bulandshahr, Patiala, and Nainital. It was discovered that Bulandshahr had the highest levels of PM pollution, followed by Patiala and Nainital. Previously, three different models were used to address the three problems of forecast, interpolation, and quality analysis. However, the suggested model can now resolve all three of these issues with just one model, namely air pollutant prediction. With a variety of inputs, including pollutant concentrations, meteorological data, and readily accessible traffic information, machine learning algorithms, such as ANN, have frequently been used to forecast air quality and make predictions. for analysing the hidden characteristics of the database, which manages both space and time information Aruna Kumari N S et al. [9] have used the random forest algorithm. The suggested system offers the anticipated outcomes for air quality and the harm that air pollution causes to human health. Particulate matter is tiny particles with the size of 2.5 . When these small particles of particulate matter enter the respiratory area, they cause various diseases such as cough, asthma attacks, high blood pressure, and heart attacks. To monitor PM concentration, R. Nithya Shree et al. [10] have used various productive machine learning models such as bagged Classification, mixture discriminant analysis, and random Forest with accuracy of 0.92, 0.87, and 0.93 respectively. To overcome the accuracy limitations of the above three methods for large sample data, extreme gradient boosting is used. Chemicals or other particles in the air that are harmful to plants, animals, and people collectively constitute air pollution. To monitor the growth of air pollution in urban areas, Nandini K et al. [11] have used machine learning algorithms such as K-means for clustering, and other classifiers such as multinominal logistic regression and decision tree. According to their research, the multinominal logistic regression model has outperformed the decision tree model in terms of accuracy.

# CHAPTER-3

# PROPOSED METHODOLGY

* 1. **FLOW CHART:**

**Diagram

Description automatically generated**

* 1. **Algorithm Proposed**

## **Model Training**

Firstly, we calculated the AQI using the values of the four air pollutants (SO2, NO2, RSPM, and SPM) using the AQI calculation formula (1), and then, we included these calculated values of AQI in the dataset to label the dataset. After that, we divided the dataset into two parts: the testing dataset and the training dataset. Using the training dataset, we train our algorithm and after the training of the model, we will test our model with the testing dataset.

**1. Using Linear Regression [19]**: Linear Regression is a supervised machine learning algorithm that is used for predictive analysis. The linear regression technique is used to make predictions on continuous or numeric variables. Linear regression algorithm shows a linear relationship between a dependent and one or more independent variables, i.e., linear regression is used to find how the value of the dependent variable changes according to the value of the independent variable.

### **2. Using Classification Algorithms:** Following the guidelines of the Ministry of Environment, Forest, and Climatic change, we have classified the dataset based on their calculated AQI values to categorize the air quality into six categories, which are, good, satisfactory, moderate, poor, very poor, and severe. This new category field is also added to our dataset, so that we can take new data as input and based on these categories, produce the corresponding result, i.e., to which category does the air quality belongs. We have used different classification algorithms such as Random Forest, Decision Tree, and Logistic Regression.

1. *Logistic Regression [18]*: Logistic Regression is a supervised machine learning algorithm that is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression uses sigmoid function to calculate the output. As an output, it gives the probabilistic values which lie between 0 and 1.
2. *Decision Tree:* The Decision Tree is a supervised machine-learning algorithm that is used to solve both regression and classification problems. The Decision Tree is a graphical representation of getting all the possible outcomes by following certain branches, i.e., by taking decisions based on the conditions.
3. *Random Forest:* Random Forest is a supervised machine learning algorithm that is a classifier that contains multiple numbers of decision trees. Random forest takes less training time as compared to other algorithms. It predicts the output of higher accuracy, even for a large dataset it produces accurate results, even if the dataset contains a large portion of the missing values.

**CHAPTER-4**

**TECHNOLOGIES TO BE USED**

* Machine Learning Algorithms such as Linear Regression, Logistic Regression, Decision Tree and Random Forest.
* HTML
* CSS
* Bootstrap

**CHAPTER-5**

**CONCLUSION**

From the analysis of the air pollution trends in the past years, we have found that Delhi, Bihar, Punjab, and Uttar Pradesh are the most polluted states in India. We have used four different algorithms linear regression, logistic regression, decision tree, and random forest. Linear regression algorithm is used for predicting the AQI values based on different values of the four pollutants, SO2, NO2, RSPM, and SPM. Classification algorithms like Logistic Regression, Decision Tree, and Random Forest are used to classify air quality into six different categories, good, satisfactory, moderate, poor, very poor, and severe. Different supervised machine learning algorithms such as Linear Regression, Logistic Regression, Decision Tree and Random Forest are used for model training and determination of the air quality in future years in different cities of India. The accuracy of different classification algorithms such as Logistic Regression, Decision Tree and Random Forest are also calculated to determine the most accurate algorithm for the prediction purpose.

**REFERENCES**

1. Mijakovski, Vladimir, and Monika Lutovska. "Monitoring of Air Pollution in Skopje in order to Determine Possible Causes." *2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*. IEEE, 2020.
2. Madhuri, V. M., GH Samyama Gunjal, and Savitha Kamalapurkar. "Air pollution prediction using machine learning supervised learning approach." *International Journal of Scientific and Technology Research* 9.4 (2020): 118-123.
3. Brauer, Michael, et al. "Taking a stand against air pollution—The impact on cardiovascular disease: A joint opinion from the world heart federation, american college of cardiology, american heart association, and the european society of cardiology." *Circulation* 143.14 (2021): e800-e804.
4. Tyagi, Ashima, Latika Kharb, and Deepak Chahal. "Scrutinizing Patterns of Air Pollution in India." *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 2020.
5. Tripathi, Kshitij, and Pooja Pathak. "Deep learning techniques for air pollution." *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE, 2021.
6. Tyagi, Bhishma, et al. "Changing air pollution scenario during COVID-19: Redefining the hotspot regions over India." *Environmental Pollution* 271 (2021): 116354.
7. Chattopadhyay, Mriduchhanda, et al. "Subjective probabilistic expectations, household air pollution, and health: Evidence from cooking fuel use patterns in West Bengal, India." *Resource and Energy Economics* 66 (2021): 101262.
8. Agrawal, Girish, Dinesh Mohan, and Hifzur Rahman. "Ambient air pollution in selected small cities in India: Observed trends and future challenges." *IATSS Research* 45.1 (2021): 19-30.
9. Kumari, NS Aruna, et al. "Prediction of air quality in industrial area." *2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. IEEE, 2020.
10. Shree, R. Nithya, G. Santhiya, and R. Bhargavi. "Assessment Of Spatial Hazard And Impact Of PM10 Using Machine Learning." *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE, 2020.
11. Nandini, K., and G. Fathima. "Urban Air Quality Analysis and Prediction Using Machine Learning." *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*. IEEE, 2019.
12. Maheshwari, Khushi, and Sampada Lamba. "Air quality prediction using supervised regression model." *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. Vol. 1. IEEE, 2019.
13. Kumar, K., and B. P. Pande. "Air pollution prediction with machine learning: a case study of Indian cities." *International Journal of Environmental Science and Technology* (2022): 1-16.
14. Mijakovski, Vladimir, and Monika Lutovska. "Monitoring of Air Pollution in Skopje in order to Determine Possible Causes." *2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*. IEEE, 2020.
15. Manisalidis, Ioannis, et al. "Environmental and health impacts of air pollution: a review." *Frontiers in public health* (2020): 14.
16. Fuller, Richard, et al. "Pollution and health: a progress update." *The Lancet Planetary Health* (2022).
17. Pandey, Anamika, et al. "Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019." *The Lancet Planetary Health* 5.1 (2021): e25-e38.
18. <https://mathworld.wolfram.com/SigmoidFunction.html> (Dated on: 17 Nov. 2022)
19. <https://www.geeksforgeeks.org/ml-linear-regression> (Dated on: 16 Nov. 2022)