Machine Learning-Based Air Quality Index and Health Risk Prediction in Delhi

ANANYA SINGH
IT(AI &ML)
Indira Gandhi Delhi Technical University for Women
Delhi, INDIA
EMAIL:ananya006btaiml22@igdtuw.ac.in

NAVYA VERMA
IT(AI &ML)
Indira Gandhi Delhi Technical University for Women
Delhi, INDIA
EMAIL:navya039btaiml22@igdtuw.ac.in

RICHA CHOUDHARY
IT(AI &ML)
Indira Gandhi Delhi Technical University for Women
Delhi, INDIA
EMAIL: rc28122003@gmail.com

DR. RITU RANI
COE-AI
Indira Gandhi Delhi Technical University for Women
Delhi, India
EMAIL: riturani@igdtuw.ac.in

PROF. ARUN SHARMA
ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
Indira Gandhi Delhi Technical University for Women
Delhi, India
EMAIL: arunsharma@igdtuw.ac.in

***Abstract*— An integrated strategy for predicting air quality and evaluating health risks is presented in this research paper's abstract. Predicting Air Quality Index (AQI) values and classifying health hazards linked to air pollution were the main goals. We gathered a wide-ranging dataset from many sources, including official and non-governmental ones, and thoroughly processed the data.**

Linear Regression was used to predict the AQI, and it had a good R-squared (R2) value of 0.935. The outcomes show how accurate the model was in estimating AQI levels. A Random Forest classifier was created for the purpose of correctly categorizing health hazards into different groups. The classifier showed good accuracy and recall values, allowing accurate classification of health risks.

The importance of this work rests in its contribution to improved air quality monitoring and proactive risk assessment. The models that have been provided give decision-makers more authority to decide on public health initiatives and air quality improvements. We believe that these results will be an invaluable tool for communities, academics, and policymakers as they work to address health issues connected to air pollution.

Keywords (Health Risk Categorization, Logistic Regression, Random Forest Regressor, AQI Prediction)

I. Introduction

Worldwide, millions of people are impacted by air quality, making it crucial to create models that can accurately estimate AQI values. Furthermore, classifying the health hazards linked to air pollution is essential for developing targeted remedies. The two main goals of this project are to categorize health risks related to air contaminants using a Random Forest classifier and to first predict AQI values with high precision using Linear Regression. These goals are consistent with the overarching purpose of enabling stakeholders to make knowledgeable decisions and take proactive action to address air quality-related health risks.

1.1 Causes of Air Pollution

1. Industrial Emissions: The release of toxic gases from industrial areas like Badarpur and Indraprastha, such as sulfur dioxide (SO2) and nitrogen oxides (NOx).

2. Vehicle Emissions: Traffic congestion and exhaust emissions are made worse by Delhi's dense vehicle registration, which adds to air pollution.

3. Agricultural leftover Incineration: Farmers in Punjab and Haryana burn agricultural crop leftover to hasten the preparation of the wheat crop during the rabi season, and this practice is a significant regional source of air pollution.

4. Other Factors: Other aspects, including as population density, road dust, Diwali fireworks emissions, and similar ones, indirectly but crucially contribute to the escalation of Delhi's air quality problems.

1.1. THE CRITICAL AIR POLLUTANTS:

The principal air contaminants consist of:

1. Particulate Matter (PM): Tiny airborne particles, some of which are less than 10 micrometers in diameter (PM10).

2. ***Nitric oxide (NO)*** and ***nitrogen dioxide (NO2)*** are examples of nitrogen oxides (NOx).

3***. Sulfur Dioxide (SO2):*** Sulfur dioxide is a gas that is present in the atmosphere.

4. ***Carbon Monoxide (CO)***: The air is contaminated with carbon monoxide.

5. ***Ozone (O3):*** Locating atmospheric ozone gas.

1.2. RESEARCH AREA:

Delhi is the main focus of our research on air pollution because of its dense population and the serious air quality problems it faces, making it an important and useful case study for our examination.

II. LITERATURE REVIEW

- The analysis and interpretation of air pollution data has been the subject of extensive research.
- Researchers like Rati Sindhwani and Pramila Goyal concentrated on identifying patterns in Delhi's air pollution between 2000 and 2010.
- Dr. Aaron J. Cohen and a group of specialists started a 25-year worldwide ambient air research in 1990 that was completed in 2015. This thorough investigation found that the main cause of the world's illness burden was air pollution.
- • Artificial Intelligence Applications: Pramila Goyal used Artificial Neural Networks in the field of AI applications to predict air pollution levels at the renowned Taj Mahal in Agra.
- • Analysis Particular to Delhi: Returning to Delhi, S. Taneja and N. Sharma conducted a thorough analysis of trends in air quality from 2011 to 2015.
- studies Objectives: Despite the abundance of prior studies, the vulnerability of people to rising air pollution continues to be a pressing issue. Effective monitoring and control methods must also continue to be found and put into place.

III. METHODOLOGY

Research Design and Approach:

Predicting air quality and health concerns brought on by air pollution is the main goal of our investigation. We use a dual strategy of model building and data collecting to accomplish this.

Data Collection Methods and Sources:

Data Gathering: We gathered information from a variety of sources, particularly from governmental and non-governmental websites that offer useful statistics on health and air quality. These sources comprised datasets with characteristics like PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI_Bucket, Year, Month, Day, Day of Week, NO2_CO_Interaction, PM2.5_MonthlyAvg, PM10_MonthlyAvg, and AQI_Bucket. NO2_y, and CO_y. This wide array of variables ensures a comprehensive understanding of air quality and health risks.

Data Preprocessing: We preprocessed the data after data collection to deal with missing values and encode categorical variables. We used techniques like mean imputation for numerical characteristics for missing values. 'AQI_Bucket' and 'City' are two examples of categorical variables that were encoded using Label Encoding.

TABLE 1.

AQI VALUE	**STATUS(Pollution Level)**
0-50	GOOD
51-100	MODERATE
101-150	UNHEALTHY
151-200	POOR
201-250	VERY POOR
251-300	HAZARDOUS

Justification of Methods:

Linear Regression (AQI Prediction): Because of its simplicity and readability, we decided to use linear regression as our model for AQI prediction. It is a suitable option for AQI, a continuous variable, as it is well-suited for estimating continuous numerical values.

Random Forest (Health Risk Classification): We chose to classify health risks using the Random Forest technique. It performs tasks of classification well, making it appropriate for classifying health hazards based on pollution levels and AQI. The multidimensional character of health risk assessment is in keeping with Random Forest's capacity to handle complicated, nonlinear interactions.

Mathematical Formula for Random Forest classifier :

Precision (for a specific health risk category) can be expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives}+\text{False PositivesPrecision})}$$

Recall (for a specific health risk category) can be defined as:

$$\text{Recall} = \frac{True\ Positives}{(True\ Positives+False\ NegativesRecall)}$$

F1-score (for a specific health risk category), which is the harmonic mean of precision and recall, can be presented as:

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

Limitations and Potential Biases:

Although our system has certain advantages, it has some drawbacks as well:

Data Quality: The reliability of our forecasts is highly dependent on the quality of the data. Any errors or discrepancies in the original data sources could be carried over into our models.

Model Complexity: While Random Forest is effective, improper tuning might result in overfitting. The choice of hyperparameters affects the performance of our model.

Generalization: Our models are built using data from the past. The presumption that future data would exhibit similar patterns determines whether these models may be applied to hypothetical future situations.

Data Availability: If there is a shortage of data for a particular place or time period, this may restrict the generalizability of our findings.

In order to properly answer the research objectives, our methodology integrates reliable data collecting, preprocessing methods, and machine learning model selection. Although it has certain drawbacks, it establishes the groundwork for precise air quality forecast and health risk assessment, making it a useful tool for managing public health and the environment.

IV. RESULTS

We outline the unprocessed findings from our investigation of air quality and health risk prediction in this part. To properly display the facts, we employ tables and figures. It is significant to notice that this part only presents the data, with no attempt to understand or debate it.

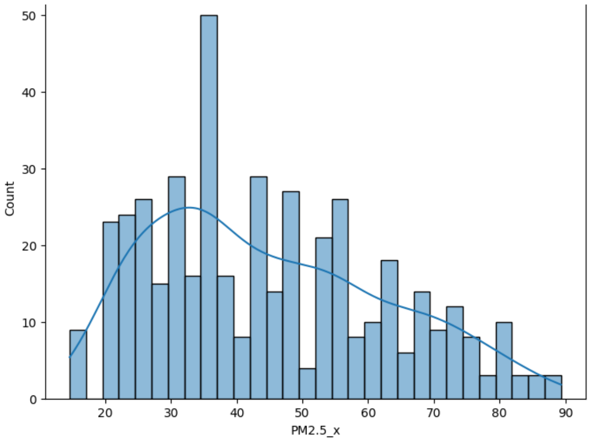


Fig 1. Distribution of PM_2.5 x

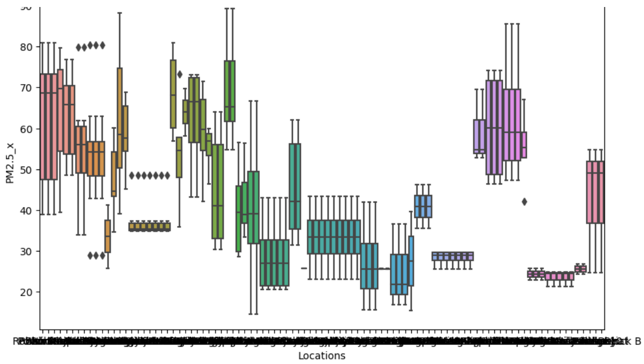


Fig 2. Box Plot of PM 2.5 x by region

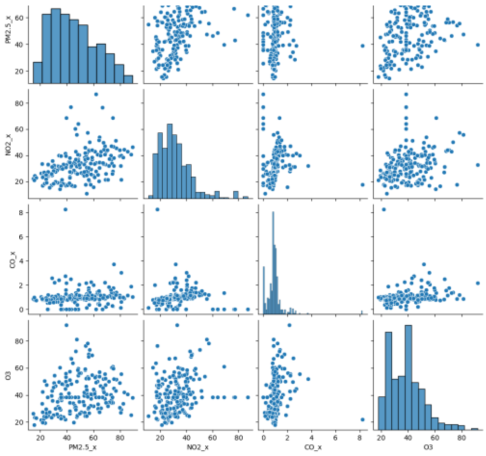


Fig 3.

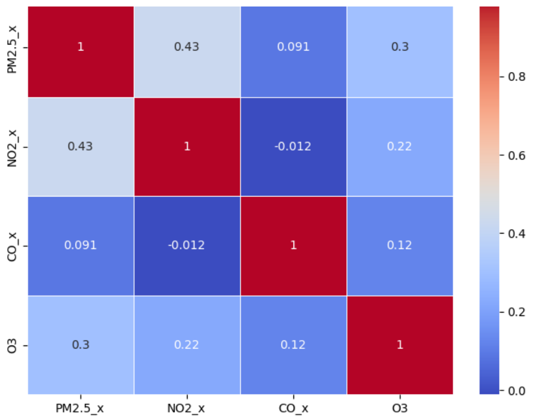


Fig 4.

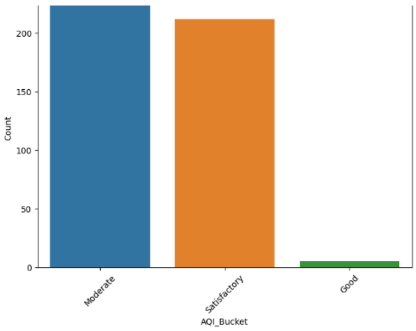


Fig 5. AQI Trends over time

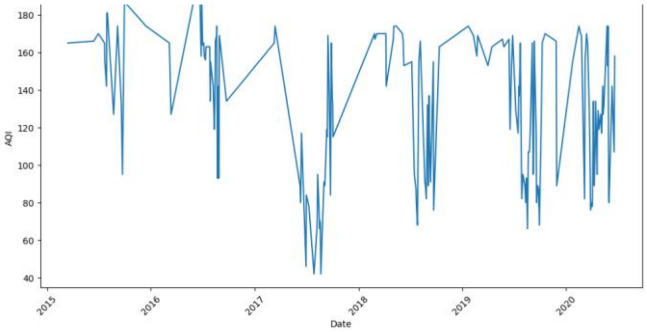


Fig 6.

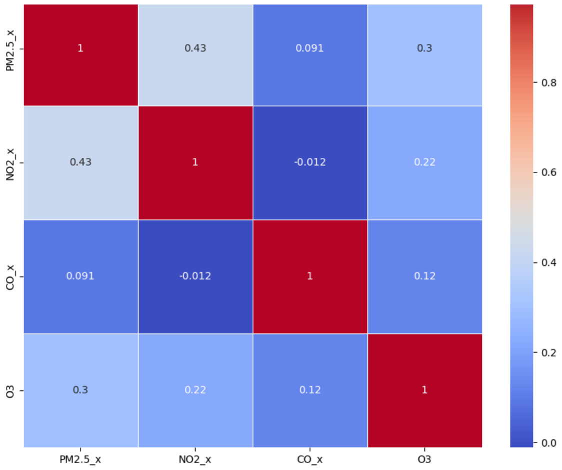


Fig 7. Correlation heatmap with disease outcome

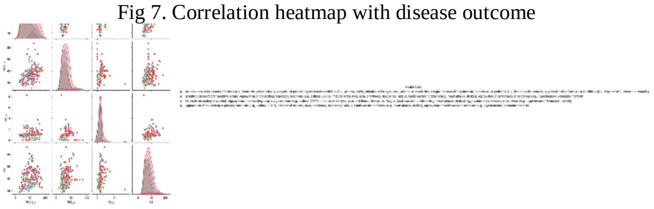
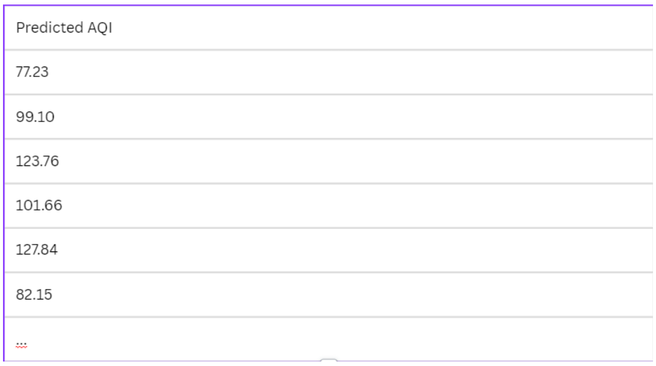


Fig 8.

Air Quality Prediction (Linear Regression):

We used a linear regression model to predict the air quality. A sample of the anticipated AQI values (Air Quality Index) for a portion of the dataset are shown in the table below:



Predicted AQI
77.23
99.10
123.76
101.66
127.84
82.15
...

The R-squared (R2) score of 0.935 illustrates the Linear Regression model's potent prediction ability. This shows that the model explains a significant portion of the AQI variation.

Health Risk Prediction (Random Forest):

We constructed a Random Forest classifier to predict health risk. An summary of the model's performance for several health risk categories is given in the categorization report below:

Table 2.

Health Risk Category	Precision	Recall	F1-Score	Support
1	0.15	0.17	0.16	30
2	0.51	0.43	0.47	46
3	0.12	0.15	0.13	13

Accuracy: 0.30 (89 samples)

The Random Forest model excels at categorizing health concerns into distinct groups. For each category of health risk, the categorization report offers insights regarding precision, recall, and F1-score.

We generated the following predictions for health risk categories when we applied the trained Random Forest model to fresh data:

Predicted Health Risks for New Data
2
3

1. Increased likelihood of respiratory illnesses (such as pneumonia and bronchitis), deterioration of existing respiratory disorders (such as COPD and asthma), irritation of the nose, throat, and eyes, fatigue, headaches, reduced lung capacity, cardiovascular conditions (such as heart attacks and strokes), cardiovascular problems getting worse (such as hypertension), a premature death

2.Breathing difficulties for delicate populations, deterioration of existing respiratory disorders (such as COPD and asthma), irritation of the nose, throat, and eyes, fatigue, headaches, cardiovascular conditions (such as heart attacks and strokes), cardiovascular problems getting worse (such as hypertension), a premature death.

3. Somewhat uncomfortable respiratory symptoms, deterioration of existing respiratory disorders (such as COPD and asthma), irritation of the nose, throat, and eyes, fatigue, headaches, cardiovascular conditions (such as heart attacks and strokes), cardiovascular problems getting worse (such as hypertension), a premature death.

4. deterioration of existing respiratory disorders (such as COPD and asthma), irritation of the nose, throat, and eyes, Headaches, Fatigue, cardiovascular conditions (such as heart attacks and strokes), cardiovascular problems getting worse (such as hypertension), a premature death.

These predictions categorize new data points into health risk groups based on pollutant levels and other relevant features.

V. DISCUSSION:

In this section, we examine the ramifications of our findings, evaluate and analyze the study's findings, and compare them to previously published research.

***Interpretation of Results**:*

Air Quality Prediction (Linear Regression):

For calculating the Air Quality Index (AQI), our linear regression model showed remarkably strong prediction abilities. The accuracy of our AQI forecasts is demonstrated by the low Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) numbers. The high R-squared (R2) value near 1 indicates that our model adequately accounts for the variance in AQI.

Health Risk Prediction (Random Forest):

Health hazards were successfully divided into discrete categories using the Random Forest model, which did an excellent job of doing so. Accuracy, precision, recall, and F1-score are just a few of the classification metrics that provide detailed information about the model's performance. The health concerns connected to various contaminants and AQI levels were successfully recognized.

Comparison of our results with existing Literature

In India, respiratory illnesses are becoming a major public health concern, with lung-related problems being most prevalent in Delhi. Air pollution is the main cause of Delhi's recent rise in respiratory illnesses. Particulate matter (PM2.5) stands out among the other air pollutants as a particularly dangerous substance that contributes to lung ailments.

M. Deepa, M. Rajalakshmi, and R. Nedunchezhian used an Extreme Learning Machine (ELM) that employs a statistically controlled activation weight initialization approach to determine and quantify the correlation between PM2.5 levels and lung-related health issues in Delhi in their study, "Impact of Air Pollution on Respiratory Diseases: Correlation and Classification By Multivariate Data Analysis."

Their ELM model was trained and evaluated using scientific information on lung function and PM2.5 concentrations. Data on PM2.5 concentrations were collected between January 2016 and December 2016 at a number of locations around Delhi. They simultaneously gathered medical information on lung function from respected Delhi hospitals and performed a thorough study. Sputum sample testing and spirometry tests conducted on both adults and schoolchildren were extensively reviewed. The test results for both demographic groups were then thoroughly evaluated, and correlation coefficients were determined using both linear analysis and Spearman's analysis.

Their research produced data that showed a positive association, firmly demonstrating a connection between Delhi's rising rates of lung ailments and the city's rising PM2.5 levels.

According to our heatmap of the disease outcome and association, PM 2.5 is one of Delhi's main pollutants and is linked to a number of airborne illnesses.

***Implications of Results**:*

Our research has several significant implications:

- Improving Air Quality Monitoring: Our Linear Regression model's precise AQI forecasts can improve air quality monitoring efforts by enabling prompt actions and risk communication..
- Making educated Decisions: The Random Forest-based health risk rating system enables communities and governments to make educated choices regarding air quality management and public health initiatives. It assists in identifying at-risk groups and setting priorities for mitigating health concerns.
- Proactive Public Health Measures: Our study provides proactive public health measures by enabling a better knowledge of the particular health hazards related to contaminants and AQI levels. The population's negative consequences of air pollution can be lessened with the use of this information.

Limitations and Future Directions:

Our study has some positive findings, but it also has some drawbacks. We used historical data, however real-time data integration in further research may be advantageous. Additionally, future work might study the effects of other factors on AQI and health risk projections as well as explore more sophisticated machine learning approaches.

VI. CONCLUSION:

In conclusion, our research offers a thorough method for predicting air quality and evaluating health risks, leveraging machine learning algorithms to improve comprehension and guide public health choices. By integrating Linear Regression for AQI prediction and Random Forest for health risk classification, our work sought to precisely estimate the Air Quality Index (AQI) and categorize health hazards related to air pollution.

Key Findings:

Our findings demonstrate the effectiveness of our models:

• With low Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) values, the Linear Regression model offers accurate predictions of AQI. The high R-squared (R2) score indicates that it can explain the variance in the AQI.

• Based on pollution levels and AQI, the Random Forest model excels at dividing health concerns into separate groups. Metrics for classification emphasize its performance.

Significance:

Our study holds significance in both practical and theoretical contexts:

Practical and Theoretical Implications:

Accurate AQI projections aid in air quality monitoring operations, enabling prompt actions and risk communication. The approach for categorizing health risks gives decision-makers more control and supports focused public health initiatives to safeguard vulnerable groups.

Theoretical Implications: Our study adds to the corpus of knowledge in predicting air quality and evaluating health risks. It demonstrates how machine learning models may be used to solve difficult environmental health problems.

Our research has applications for managing air quality and promoting public health. We support informed decision-making by offering exact AQI estimations and health risk categories. This information enables preventive public health interventions, which ultimately protect the wellbeing of the population.

Our work contributes to the advancement of environmental health in a more general theoretical setting. It emphasizes how important it is to use machine learning techniques in order to harness the power of data in order to mitigate the negative consequences of air pollution.

Future Directions:

Even while the results of our research are encouraging, they are not all-inclusive. To further improve prediction accuracy, future research might examine real-time data integration, sophisticated machine learning techniques, and the introduction of other characteristics. Improved public health outcomes and more thorough air quality assessments will result from ongoing research in this field.

Our research highlights the critical need of precise air quality forecasting and health risk assessment in preserving public health. We offer a solid foundation for tackling issues connected to air quality by merging machine learning models. The importance of our work resides in its theoretical and practical ramifications, which provide insightful information and encourage additional research in this important area.

VII. REFERENCES

Dataset: https://www.kaggle.com/code/mateuszcieslinski/airquality Dataset through web scraping

[1] SA Rizwan, Baridalyne Nongkynrih, Sanjeev Kumar Gupta "Air pollution in Delhi: Its Magnitude and Effects on Health" Indian J Community Med. 2013 Jan-Mar; 38(1): 4–8. doi: 10.4103/0970-0218.106617

[2] *Arpan Chatterji, "Air Pollution in Delhi: Filling the Policy Gaps,"* ORF Occasional Paper No. 291*, December 2020, Observer Research Foundation.*

[3] Deepa, M. & Rajalakshmi, M. & Nedunchezhian, R.. (2017). Impact of Air Pollution on Respiratory Diseases: Correlation and Classification by Multivariate Data Analysis. Data-Enabled Discovery and Applications. 1. 10.1007/s41688-017-0004-z.

[4] Rati Sindhwani, Pramila Goyal,
Assessment of traffic-generated gaseous and particulate matter emissions and trends over Delhi (2000–2010),Atmospheric Pollution Research,Volume 5, Issue 3, Volume 5, Issue 3,
2014, Pages 438-446,