Air Quality Index Prediction Using Machine Learning

Abhishek Mourya[1]
Student, Department of MSc. IT,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
mouryaa201@gmail.com

Dr. Pallavi Devendra Tawde[2]
Assistant professor,
Department of IT and CS,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
pallavi.tawde09@gmail.com

ABSTRACT: Air pollution poses a significant threat to public health, particularly in rapidly developing countries like India, where urbanization and industrialization contribute to diverse sources of pollutants. This research employs machine learning techniques to comprehensively analyse and predict air quality levels across different regions of India. Extensive air quality data, encompassing various pollutants and temporal variations, is collected and compiled. State-ofthe-art machine learning models are developed to identify key influencing factors, including industrial emissions, vehicular traffic, agricultural practices, and meteorological conditions. The models are evaluated for accuracy and reliability in predicting air quality levels. The study aims to provide valuable insights into the complex dynamics of air pollution in the Indian context. By understanding the contributing factors and predicting future air quality levels, this research contributes to the development of targeted and effective air quality management strategies.

Keywords: Machine Learning, Air Pollution, AQI.

1. Introduction

Air quality analysis and prediction play a pivotal role in safeguarding public health and the environment. As urbanization and industrialization continue to rise, the impact on air quality becomes a matter of global concern. Poor air quality is associated with various health issues, including respiratory and cardiovascular diseases, making it crucial to monitor and predict air quality levels.

India faces unique challenges regarding air quality due to a combination of rapid industrial growth, urbanization, and diverse geographical and meteorological conditions. The country grapples with high levels of air pollution from vehicular emissions, industrial activities, and agricultural practices. The seasonal variations, such as the winter smog in northern regions and monsoon-related pollution in other parts, contribute to the complexity of the issue. Understanding and addressing these challenges are critical for developing effective air quality management strategies tailored to the specific needs of the Indian context.

The primary contributors to air pollution include Ozone (O3), Nitrogen dioxide (NO2), Carbon Monoxide (CO), Sulphuric oxide (SO2), and Particulate Matter (PM). These gases, though invisible and imperceptible, are produced from various sources such as the burning of fossil fuels, wood combustion, industrial boilers, and volcanic eruptions. Despite their elusive nature, these pollutants have profound effects on human health. Prolonged exposure to these gases is linked to serious health issues, including cancer, birth defects, and respiratory problems. It is imperative to address and mitigate the sources of these pollutants to safeguard human well-being and environmental health.

2. Review of Literature

To understand and summarize how well machine learning helps in forecasting the AQI. To identify the best supervised machine learning algorithms to predict the AQI. We performed literature review and collected some previous research papers that helped us to achieve the above. The research papers were presented below.

No	Title	Observation
[1]	Okokpujie et al. (2018): "A Comprehensive Air Pollution Monitoring Framework"	Presents a detailed framework for monitoring air pollution. Contributes to the field of civil engineering and technology.
[2]	Veljanovska, K., & Dimoski, A. (2018): "Air Quality Index Prediction Using Simple Machine Learning Algorithms"	Explores the prediction of AQI using straightforward machine learning algorithms.
[3]	Zhu, D et al. (2018): "A Machine Learning Approach for Air Quality Index Prediction"	Introduces a machine learning approach for predicting AQI. Emphasizes feature regularization and optimization.
[4]	AQI Forecasting: A. Kumar and P. Goyal (2011)	Forecasting of air quality in Delhi using principal component regression technique.
[5]	H. Li et al. (2018): "Impact of emissions on air quality"	Analyzes the influence of emissions on the air quality index using principal component regression.
[6]	Y.-C. Liang et al. (2020): "Machine learning-based prediction of air quality"	Focuses on machine learning-based prediction of air quality.

Table 1: Review of literature

3. Methodology

A. Data Set

1. Pollutant Data: The dataset utilized for training the system to detect air quality includes essential information on pollutants. Attributes such as Carbon Monoxide (CO), Sulfur Dioxide (SO2), and Ozone (O3) are incorporated to capture key pollutant levels.

2. Meteorological Data: In addition to pollutant data, the dataset encompasses meteorological information vital for a comprehensive understanding of air quality dynamics.

B. Using Regression Algorithms

1. Linear Regression: Linear Regression is a supervised learning algorithm employed for regression tasks.

2. Decision Tree: Decision Tree Regression is a versatile model used for both regression and classification problems.

3. Random Forest: Random Forest is an ensemble method that utilizes bagging techniques.

C. Using Classification Algorithms

1. **Logistic Regression:** Logistic Regression is a widely used linear model for binary classification problems.
2. **Decision Tree Classifier:** Decision Trees are non-linear models that recursively split the data based on feature values.
3. **Random Forest Classifier:** Random Forest Classifier is an outfit learning strategy that builds numerous choice trees and consolidates their forecasts.
4. **K-Nearest Neighbors (KNN) Classifier:** KNN could be a and natural calculation that classifies information focuses based on the majority class of their k-nearest neighbors.

4. Model with experiment result

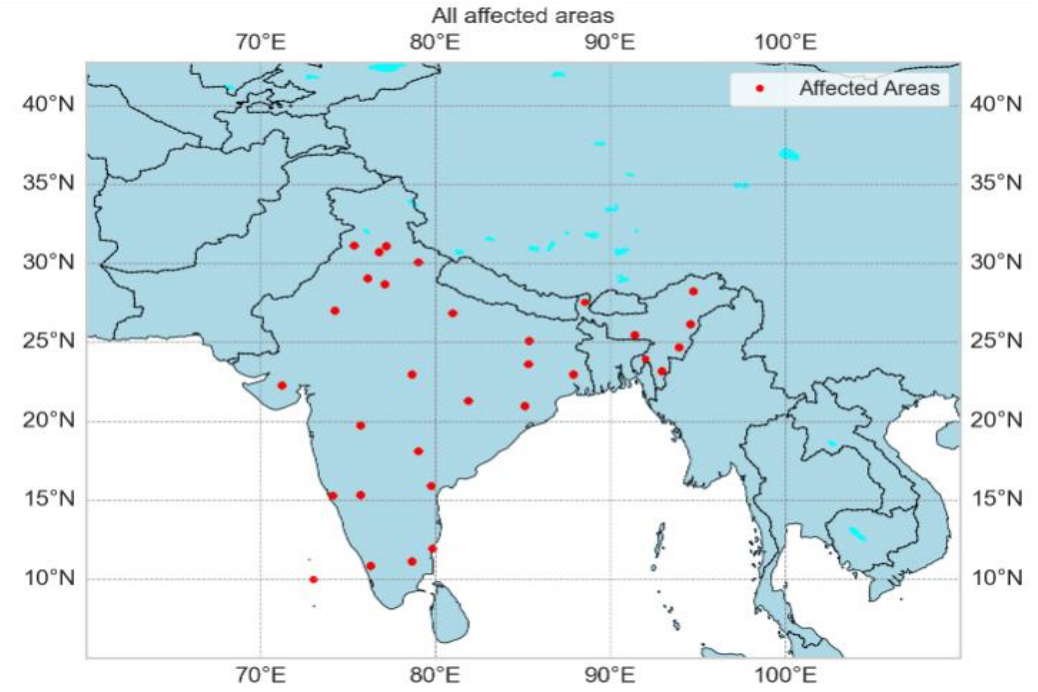

Fig 1: Plotting affected areas

Here, we have calculated the air quality index (AQI) of every data value based on the level of sulphur dioxide and nitrogen dioxide in air and it is calculated as per indian govt standards.

	sampling_date	state	si	ni	rpi	spi	AQI
0	February - M021990	Andhra Pradesh	6.000	21.750	0.0	0.0	21.750
1	February - M021990	Andhra Pradesh	3.875	8.750	0.0	0.0	8.750
2	February - M021990	Andhra Pradesh	7.750	35.625	0.0	0.0	35.625
3	March - M031990	Andhra Pradesh	7.875	18.375	0.0	0.0	18.375
4	March - M031990	Andhra Pradesh	5.875	9.375	0.0	0.0	9.375

Fig 2:

Plotting the naïve forecast approach.

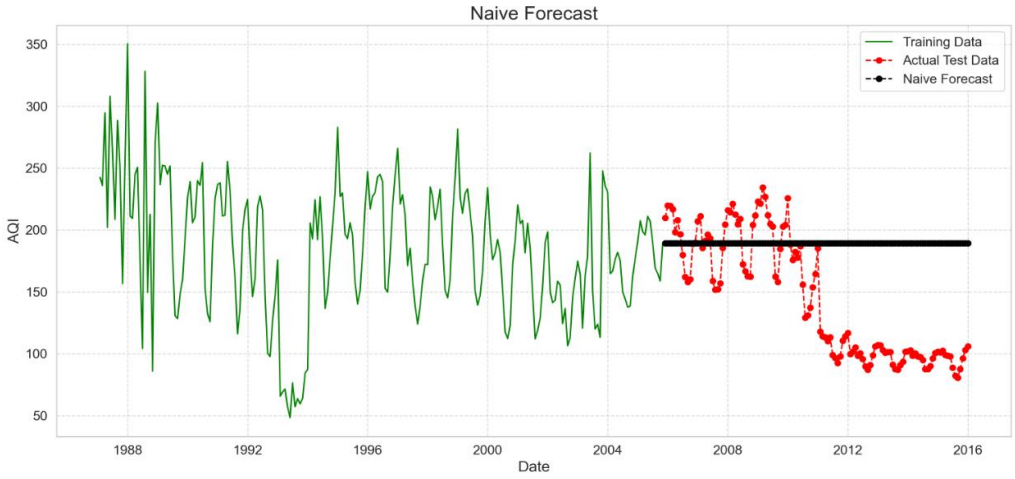

Fig 3:

Regression Algorithms:

Model	RMSE	R-Square	Accuracy
linear regression	31.91	0.9263	77.63
Decision Tree	2.12	0.9996	98.13
Random Forest	2.25	0.9996	98.41

Fig 4: Comparison of performance metrics for all models

By observing the table, when compared to all algorithms the model has lower MAE, lower RMSE and higher r-squared error when built using Decision Tree and random forest regression algorithm. This indicates that Decision Tree regression model is good for forecasting the AQI.

Classification Algorithms:

Model	Accuracy	KappaScore
logistic regression	58.7	0.301
Decision Tree	99.9	0.999
Random Forest	99.9	0.999
K-Nearest Neighbours	99.2	0.988

Fig 5: Comparison of performance metrics for all models

By observing the table, when compared to all algorithms Both the Random Forest and Decision Tree models are performing exceptionally well, with the same accuracy and KappaScore. If the goal is to maximize predictive performance, the Random Forest model may be the preferred choice.

5. Conclusion

In this research study, we aimed to assess and predict air quality variations, specifically focusing on the Air Quality Index (AQI) across India. We applied various data preprocessing techniques and implemented a Naive Forecast approach to establish a baseline for comparison with more advanced models.

The data were processed to calculate individual pollutant indices, including Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Respirable Particulate Matter (RPM), and Suspended Particulate Matter (SPM), adhering to Indian government standards. Additionally, the AQI was calculated based on these sub-indices to provide a comprehensive measure of overall air quality.

The Naive Forecast approach, depicted in Figure, serves as an initial benchmark for predicting AQI variations over time. This approach assumes that future AQI values are equal to the last observed value in the training set. The plot illustrates the actual AQI values from the training and test sets alongside the predicted values using the Naive Forecast.

Moving forward, more sophisticated models, including Lasso Regression, Ridge Regression, and Time Series analysis, will be explored to enhance the accuracy of AQI predictions. These models will undergo further refinement and validation to determine their effectiveness in capturing complex patterns and trends in air quality data.

6. References

[1] Okokpujie et al. (2018). "A Comprehensive Air Pollution Monitoring Framework." International Journal of Civil Engineering and Technology (IJCIET). This research paper presents a detailed framework for monitoring air pollution, aiming to contribute to the field of civil engineering and technology.

[2] Veljanovska, K., & Dimoski, A. (2018). "Air Quality Index Prediction Using Simple Machine Learning Algorithms." International Journal of Emerging Trends & Technology in Computer Science, 7(1). The paper explores the prediction of Air Quality Index using straightforward machine learning algorithms.

[3] Zhu, D et al. (2018). "A Machine Learning Approach for Air Quality Index Prediction: Feature Regularization and Optimization." Big Data and Cognitive Computing, 2(1), 5. The research introduces a machine learning approach for predicting the Air Quality Index, emphasizing feature regularization and optimization.

[4] A. Kumar and P. Goyal, "Forecasting of air quality in Delhi using principal component regression technique," Air Pollution Research, vol. 2, no. 4, pp. 436–444, 2011. This study focuses on forecasting air quality in Delhi using the principal component regression technique.

[5] C. Li et al. "Research on air quality prediction based on machine learning," in 2021 2nd International Conference on Human-Computer Interaction (ICHCI), 2021, pp. 77–81. The paper delves into air quality prediction using machine learning, presented at the International Conference on Human-Computer Interaction.

[6] H. Li et al. "Analyzing the impact of emissions on air quality index based on principal component regression," Journal of Cleaner Production, vol. 171, pp. 1577–1592, 2018. This research analyzes the influence of emissions on the air quality index using principal component regression.

[7] Y.-C. Liang et al. "Machine learning-based prediction of air quality," Applied Sciences, vol. 10, no. 24, p. 9151, 2020. The study focuses on machine learning-based prediction of air quality, published in the journal Applied Sciences.

[8] M. K. Tiwari et al. "Urban air quality modeling based on multi-dimensional collaborative support vector regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang," in 2017.

[9] Pallavi Devendra Tawde, Dr. Sarika Chouhan (2019), "A Survey of Machine Learning Techniques for Student Performance and Placement" Journal of Xi'an University of Architecture & Technology, Vol. XI, Issue XII, 2019, ISSN 1006-7930.

[10] Kaggle - Indian Air Quality Analysis Datasets. The Kaggle dataset provides information for Indian air quality analysis, contributing to the broader research on this topic.