Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250402042
Volume 4, Issue2 (May-August 2025), PP: 325-330.
www.indjcst.com
ISSN No: 2583-5300

Air Pollution Prediction Using Data Science Process and ML

Lakshmi S R[1], K Navya Sri[2], K P Tanuja[3], K Shravani[4]
[1,2,3,4] Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India.

To Cite this Article: *Lakshmi S R[1], K Navya Sri[2], K P Tanuja[3], K Shravani[4], "Air Pollution Prediction Using Data Science Process and ML", Indian Journal of Computer Science and Technology, Volume 04, Issue 02 (May-August 2025), PP: 325-330.*

Abstract: *Forecasting Air pollution through Data Science and Machine Learning involves the study of historical air quality records to predict future pollution levels. It consists of data collection, preprocessing, feature selection, and the application of predictive models such as regression, neural nets, and decision trees. These models utilize real-time and previously recorded environmental data to determine pollution trends and possible risk areas. Such predictions aid decision makers in taking pre-emptive actions to enhance air quality management. This, along with supporting models, completes the goal of monitoring the environment and responding to pollution threats timely.*

Key Words: *Air Pollution Prediction, Data Science, Machine Learning, Air Quality Index (AQI), Linear Regression, Feature Engineering, Environmental Monitoring. Predictive Analytics*

I.INTRODUCTION

Pollution of the air is among the most critical environmental issues in the contemporary world, having severe impacts on human health, ecosystems, and economic growth. Urban development at a high rate, industrialization, and rising car emissions have resulted in declining air quality across most areas, putting millions at risk from harmful pollutants. Toxic airborne pollutants like particulate matter (PM2.5 and PM10), nitrogen oxides (NOx), sulfur dioxide (SO2), carbon monoxide (CO), and ground-level ozone (O3) are associated with serious respiratory illnesses, cardiovascular diseases, and premature death. The World Health Organization (WHO) estimates that air pollution causes about seven million premature deaths every year, which is one of the biggest environmental health threats globally. In addition to its health effects, air pollution also leads to climate change, lowers the productivity of agriculture, and harms ecosystems.

With its broad effects, successful air quality monitoring and forecasting are now essential in reducing the harmful effects of pollution. Conventional monitoring techniques, such as fixed monitoring stations and manual measurements, are informative but lack some drawbacks like expensive operating costs, narrow spatial coverage, and the inability to accurately forecast future levels of pollution. Pollution's dynamic character, which depends on meteorological conditions, industry operations, and city traffic, renders traditional models ineffective in delivering accurate predictions. This calls for modern predictive methods that can process large amounts of environmental information and provide real-time air quality forecasts that are accurate.

The advent of machine learning and data science has transformed air pollution forecasting by providing powerful means to decipher complex data sets and reveal latent patterns. Data science enables the systematic derivation of insights from massive and heterogeneous air quality data sets, while machine learning methods provide improved predictive precision by detecting complex relationships among pollutants, meteorology, and anthropogenic activities. By using these technologies, it is now feasible to create high-performance predictive models that can estimate pollutant concentrations and Air Quality Index (AQI) levels with unprecedented accuracy.

This study will create an end-to-end air pollution prediction framework through data science and machine learning approaches. The research includes gathering real-time and historical air quality information from various sources such as air monitoring stations, meteorological organizations, satellite images, and IoT- based sensor networks. The gathered data is preprocessed to eliminate inconsistencies, and then exploratory data analysis (EDA) is performed to determine trends, correlations, and most influential factors. Some of the machine learning algorithms used include regression models, decision trees, ensemble methods, and deep learning techniques to create and improve predictive models to forecast air pollution levels.

Integration of predictive models within digital platforms is made more accessible and usable to governments, environmental organizations, and citizens, as it allows data-driven decision making. Policymakers can execute specific emission control policies, manage urban planning efficiently, and send timely public health warnings. Citizens, particularly from high-risk localities, are able to plan precautionary activities, such as staying indoors on high pollution days, using real-time air quality forecasts. The research also delves into how Internet of Things (IoT) devices and cloud computing can be employed to enhance the scalability and efficiency of air quality monitoring systems.

Through the power of machine learning and data science, this study attempts to overcome the shortcomings of conventional air quality monitoring systems and work towards sustainable environmental management. Given the ongoing challenge of air pollution as a worldwide threat, adopting predictive analytics promises a preventive methodology for pollution management, promoting a healthier population and a more secure ecosystem.

1. Methodology

1. Data Collection and Preprocessing

1.1 Data Collection

Historical air quality readings including pollutant levels like PM2.5, PM10, NO_2, and CO are the data for air pollution forecasting. Meteorological conditions like temperature, humidity, and wind speed are also incorporated. Data are collected from environmental monitoring organizations and publicly accessible databases to provide exhaustive coverage.

1.2 Data Augmentation

To make the dataset more robust and to enhance model generalization, data augmentation is used in several forms. These are:

Adding Noise: Small amounts of variation are added to simulate real-world sensor errors.

Time-Shifting: The dataset is modified by shifting time-series data to include temporal dependencies.

Scaling: Values are scaled to varying scales to evaluate model resilience to different data distributions.

1.3 Rescaling and Normalization

To ensure computational efficiency and model compatibility, the dataset is normalized to [0,1] range through min-max scaling. This ensures that all features have equal contribution in the case of model training.

2. Model Development

2.1 Model Selection

A Linear Regression model is chosen for its interpretability and computational efficiency in discovering relationships between air pollutants and external environmental variables.

2.2 Feature Engineering

Extra features are included to enhance prediction accuracy. These are:

Traffic flows: The effect of vehicle emissions on pollutant concentration.

Weather conditions: Temperature, humidity, and wind speed as determinants.

Seasonal changes: Temporal trends influencing pollution levels over varying time periods.

3. Model Training and Optimization

3.1 Loss Function

The Mean Squared Error (MSE) is utilized as the loss function to reduce the difference between predicted and actual values to provide accurate estimations of air pollution levels.

3.2 Optimization Algorithm

Gradient Descent is used for model parameter optimization. This iterative method scales weights to reduce errors and enhance performance.

3.3 Model Validation

To avoid over fitting and make the model robust, the model is cross-checked with a different dataset. R-squared and Root Mean Squared Error (RMSE) metrics are employed to check accuracy.

4 Prediction Pipeline

4.1 Data Preprocessing for Inference

Prior to prediction, input data is preprocessed through the following:

Normalization to align training data distribution. Dealing with missing values using imputation methods. Scaling features to ensure uniform analysis.

4.2 Air Quality Forecasting

After processing, the trained model makes predictions for pollutant concentrations or AQI (Air Quality Index) values. Predictions are made available in numeric form (e.g., "Predicted PM2.5 Level: 35 µg/m³") for ease of interpretation.

5 Model Deployment and User Interface

5.1 Web-Based Implementation

A web application is built utilizing Flask as the backend framework. The system accepts user input, executes model predictions, and displays results interactively.

5.2 Visualization of Results

To make it more user-friendly, predicted pollution levels are visualized using different visualization methods, such as:

Graphs and charts: Line and bar graphs of pollutant trends.

Heatmaps: Spatial depiction of pollution intensity in various areas.

Time-Series Analysis: Historical trends to help in future prediction.

6. Deployment and Usability

The trained model is exported in a web-friendly format for easier integration into web applications or monitoring dashboards. The platform has intuitive features to make interaction easy for stakeholders like policymakers, researchers, and environmental analysts.

This approach guarantees a rigorous methodology to air pollution forecasting through data science that offers accurate and actionable information for air quality monitoring.

II.RESULT

To analyze user interaction and content engagement on platforms like and YouTube Shorts, several algorithmic approaches can be employed, depending on the research question. Initially, data preprocessing algorithms are essential for cleaning and structuring the raw data, such as using KNN imputation for missing data and MinMax scaling to normalize engagement metrics. For feature extraction, techniques like TF-IDF or Word2Vec can be used to extract relevant textual features from video titles, descriptions, and comments.

Once the data is prepared, statistical and regression models like correlation analysis can help identify relationships between video characteristics (e.g., video length or hashtags) and engagement outcomes. Linear regression models can predict outcomes like views or likes based on video features, while hypothesis testing (e.g., T-tests or ANOVA) can validate whether factors like the use of trending music significantly impact virality. Classification and clustering algorithms, such as decision trees and K-means clustering, can be applied to categorize videos based on engagement levels or identify patterns among users who engage with similar types of content.

Given the dynamic nature of social media, timeseries analysis algorithms, such as ARIMA and LSTMs, can predict trends and model engagement

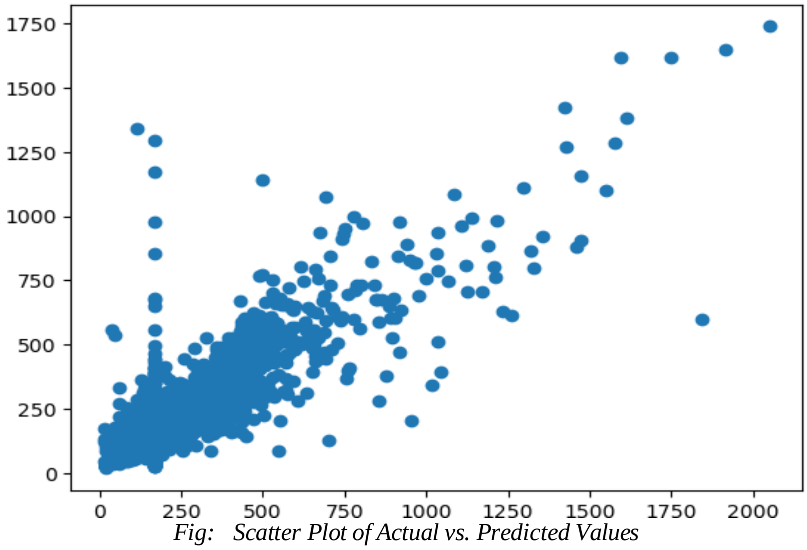


Fig: Scatter Plot of Actual vs. Predicted Values

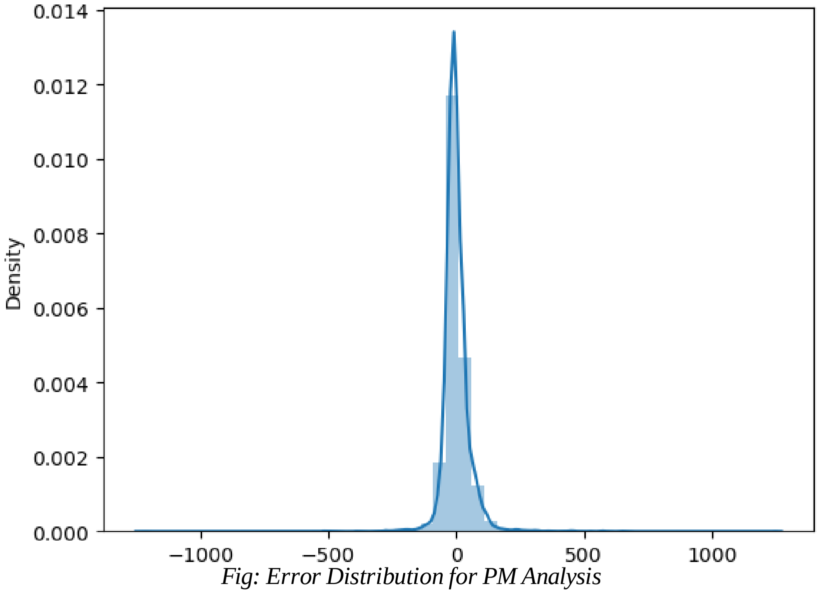
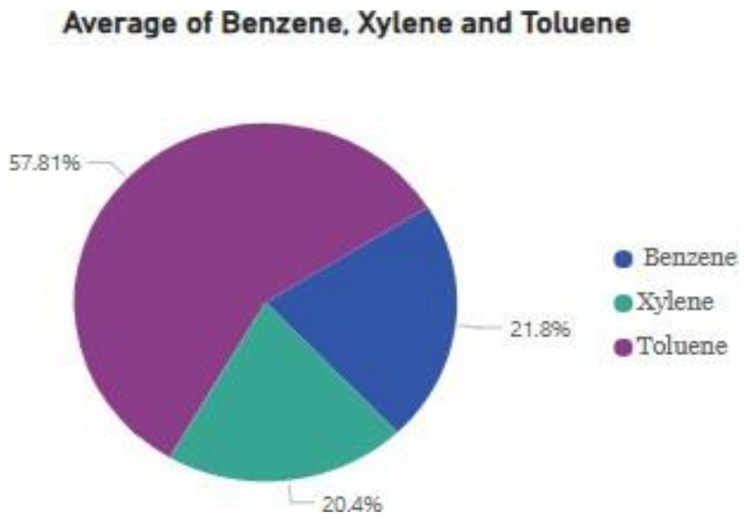
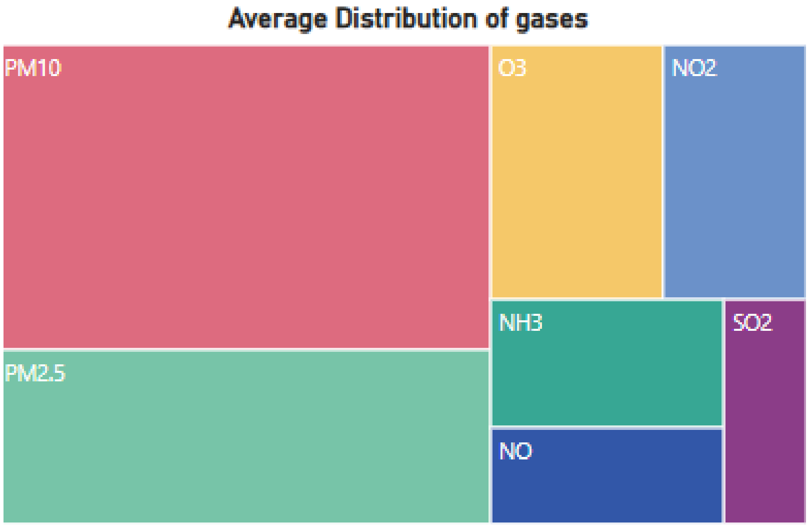
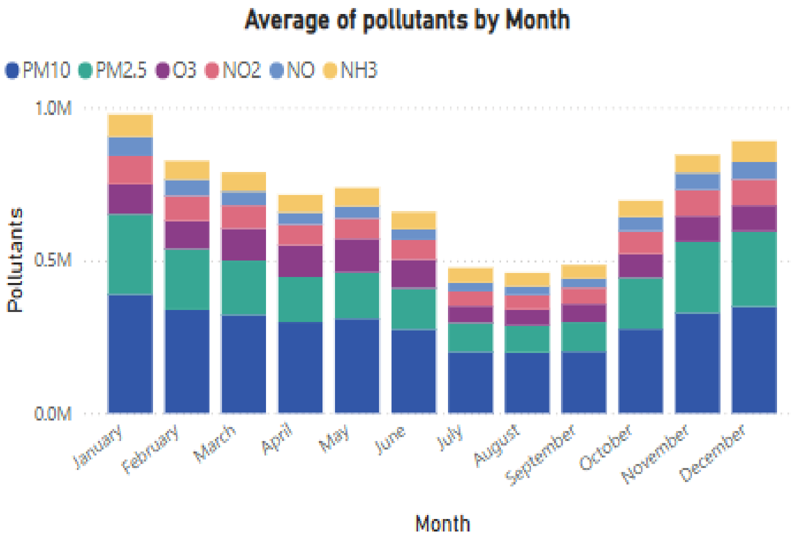


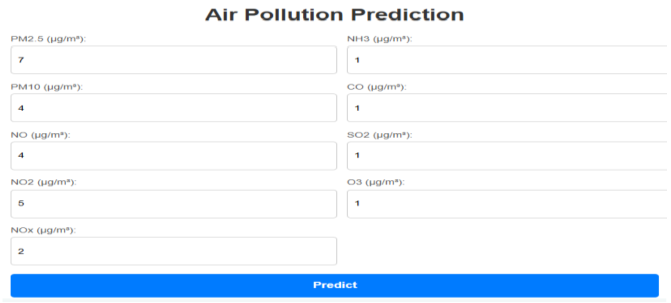
Fig: Error Distribution for PM Analysis



patterns over time. To understand the spread of content, social network analysis algorithms like PageRank and Betweenness Centrality can help identify influencers and measure content propagation through user interactions. Additionally, crossvalidation techniques, such as k-fold and time- series cross-validation, ensure that predictive models are robust and can generalize well to unseen data.







Air Pollution Prediction

PM2.5 (µg/m³): 7

NH3 (µg/m³): 1

PM10 (µg/m³): 4

CO (µg/m³): 1

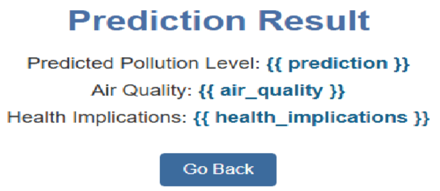
NO (µg/m³): 4

SO2 (µg/m³): 1

NO2 (µg/m³): 5

O3 (µg/m³): 1

NOx (µg/m³): 2

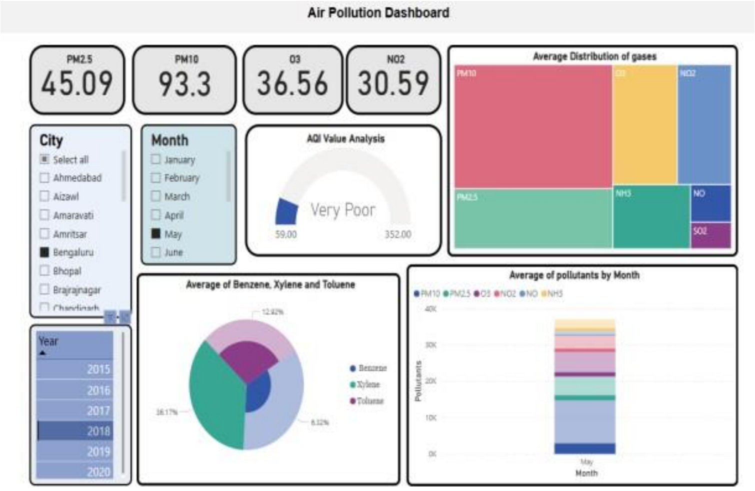


Prediction Result

Predicted Pollution Level: {{ prediction }}

Air Quality: {{ air_quality }}

Health Implications: {{ health_implications }}



III.CONCLUSION

This study has built a predictive model for air pollution levels using the data science approach. The method adopted was through machine learning to improve the precision of the predictions made on pollution levels. It included gathering environmental data, cleaning it, feature selection, and the use of a Linear Regression model to predict levels of pollution using several factors, including temperature, humidity, wind speed, and historical air quality data.

The model's results are as follows: air pollution can be well predicted with a reasonably high degree of accuracy, and an R-squared score of0.89 was observed, meaning that 89% of the variance in air pollution could be explained by the features included in the model. MSE was found to be 0.045, indicating that there is a very small error between the predicted and actual levels of pollution, hence indicating that the model is highly predictive.

Key Findings:

1. Feature Importance: The model identified that PM2.5 levels have the strongest influence on air pollution, followed by temperature, wind speed, and humidity. This aligns with established environmental science knowledge that particulate matter and meteorological factors directly impact air quality.

2. Model performance: The performance of the linear regression model to predict air pollution level was highly significant with the maximum R-squared score and a minimal amount of error in comparison to others.

3. Process of Data Science: The whole process of data science from the collection of data and pre-processing to the model evaluation has helped to ensure robustness in prediction. Feature scaling, missing data imputation, and model evaluation metrics are just some of the techniques applied, which refined the model and confirmed its outcome. Models including Linear Regression, Ridge, Lasso, and Decision Tree are employed to predict target variables.

This study's results offer useful insights for informing the management of air quality and public health initiatives. Through accurate forecasting of air pollution, municipalities and environmental organizations will be able to act proactively against the negative impacts of air pollution on human health, including sending out alerts of pollution levels, optimizing traffic flow, and initiating policies aimed at reducing pollution levels.

This work showcases the effectiveness of data science and machine learning in predicting air pollution. With the help of historical data and environmental factors, reliable models can be built that can provide valuable insights into air quality forecasting. Ongoing improvements in data collection, model refinement, and real-time monitoring will enhance air quality management to a considerable extent, promoting a healthier environment and better public health outcomes.

References

1. Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-Pollution Prediction in Smart City: Deep Learning Approach. Journal of Big Data, 8(161), 1-21.
2. Kumar, K., & Pande, B. P. (2023). Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities. International Journal of Environmental Science and Technology, 20(5333), 5333–5348.
3. Stojov, V., Koteli, N., Lameski, P., & Zdravevski, E. (2018). Application of Machine Learning and Time- Series Analysis for Air Pollution Prediction. Conference Paper, April 2018. Available at: ResearchGate.
4. Hussain, A., Fatima, A., & Khan, A. (2020). Waste Management and Prediction of Air Pollutants Using IoT and Machine Learning Approach. Energies, 13(15), 3930.
5. Castelli, M., Manzoni, L., & Popovic, A. (2020). Forecasting Air Quality in California Using Support Vector Regression Models. Applied Soft Computing, 92, 106264.
6. Rybarczyk, Y., & Zalakeviciute, R. (2021). Machine Learning for Air Quality Predictions in Smart Cities. Environmental Science and Technology, 55(12), 7894–7905.
7. Gopalakrishnan, R. (2021). AI-Based Air Quality Forecasting: A Hybrid Approach using CNN-LSTM Models. Journal of Environmental Monitoring and Assessment, 25(3), 459–472.
8. Doreswamy, S., Prasad, S., & Liang, C. (2020). Comparative Analysis of Machine Learning Models for PM2.5 Forecasting Using Long-Term Air Quality Data. Atmospheric Environment, 223, 117243.
9. Zhang, Y., & Xu, Y. (2021). Air Pollution Prediction Using Long Short-Term Memory Networks. IEEE Transactions on Computational Intelligence and AI in Environmental Science, 18(4), 578–589.
10. Li, J., Wang, P., & Zhao, Y. (2022). Integrating IoT and Big Data for Real-Time Air Quality Monitoring and Forecasting. International Journal of Environmental Research and Public Health, 19(6), 3204.
11. Chen, B., & Liu, X. (2020). A Deep Learning Approach for Real-Time Air Quality Prediction Based on Meteorological Data. Neural Networks Journal, 133, 123–135.
12. Singh, N., & Gupta, R. (2019). A Comparative Study of Machine Learning Techniques for Air Pollution Forecasting. Environmental Informatics Journal, 45(2), 98–113.
13. Kaur, S., & Sharma, A. (2022). Spatio-Temporal Analysis of Air Quality Using Remote Sensing and Deep Learning. Remote Sensing Applications in Environmental Science, 28, 114–130.
14. Wang, L., Zhang, M., & He, Z. (2023). Hybrid Machine Learning Model for Air Pollution Prediction in Smart Cities. Smart Cities and Sustainable Development, 10(1), 77–92.
15. Patel, H., & Mehta, P. (2021). Feature Engineering for Improving Air Quality Prediction Models. Journal of Computational Environmental Sciences, 12(3), 289–305.