Assignment 1: Part D
(Comprehensive Final Report)

By,

Nishant Singh a1941585

The University of Adelaide

4536_COMP_SCI_7209 Big Data Analysis and Project

Professor: Dr. Hussain Ahmad

Abstract:

Air Pollution during winters in India is driven by crop stubble burning and festival related emissions. This report evaluates machine learning models to predict AQI and determine its most important predictors (PM 2.5, PM 10). XGBoost (eXtreme Gradient Boosting) depicts the highest predictive power with effective prediction and generalization. The report also evaluates effects of a few critical events such as Covid-19 lockdown, and firecracker ban on the Air Quality. Finally, the report also highlights few important changes that government did to reduce air pollution in India.

1. Introduction

Delhi is the capital of India situated in the Northern part and Ahmedabad on the western coast of India are the highest polluted states of India in terms of Air Quality, Ahmedabad average AQI 454 and Delhi 260 which are in the Hazardous to poor category. On prolonged exposure to such hazardous air, humans can develop major medical conditions such as chronic obstructive pulmonary disease, asthma, cancer, and heart disease (World Health Organization, 2025).

1.1. Motivation

The motivation for this project arises from the need to analyze air quality pollutant values and obtain patterns in the pollution trend of India which can help drive significant change to the lives of billions of people and improve their health. By predicting and analyzing these trends can help the government provide efficient care as well as draft new policies that can help curb air pollution.

1.2. Research Questions

By answering the following question, we can help the policy makers make the correct decisions:

1. AQI predictions for New Delhi and Ahmedabad during high pollution months (October to February) using key pollutant indicators to understand the air quality variations resulting due to festivals and crop stubble burning

Additional Questions:

2. Which festival results in the worst air quality in India?
3. How did Covid-19 impact the air quality?

1.3. Contributions

1. Data Preprocessing Workflow: designed a comprehensive tidymodels workflow to handle missing values, merging and scaling of the datasets for the ML models
2. ML Model Optimization using Hyper-parameter tuning
3. Proposed Recommendations to reduce air pollution and safety instructions
4. Reproducibility: shared the source code on GitHub to allow for replication and extension of the report

2. Literature Review

Study	Region	Models Used
(Ravindiran et al., 2023)	Visakhapatnam, India	LightGBM, Random Forest, Catboost, Adaboost, and XGboost
(Mujtaba et al., 2025)	Lahore, Pakistan	SARIMA, LSTMS, seasonal autoregressive integrated moving-average with exogenous
(Pande et al., 2024)	Delhi, India	bidirectional recurrent neural networks
(Ganguli et al., 2025)	Various states of India	Deep Learning Using RNN
(Rahman et al., 2024)	Global	Random Forest, Support Vector Machines, KNN, Decision Trees

Table 1: comparison table of literature review

Compared to other past works on the air quality predictions, this report focuses on most polluted states like (Pande et al., 2024) but uses an array of ML models such as Random Forest, XGBoost, Lasso Regression. The report however does not use any deep learning algorithm and prioritizes the five of the most polluted months and studies the impact of various factors like Diwali, stubble-burning periods inspired by Pande et al.’s temporal analysis.

3. Research Methodology

The research methodology followed the steps of data science life cycle as mentioned below:

1. Data Collection: The data was collected from multiple sources such as Kaggle, Unified Portal for Agricultural Statistics (UPAg) and internet websites such as timeanddate.com

   The following datasets will be used in the report:

   a. city_hour.csv – Air Quality Data (Rao 2020) which contains the time series observation of air pollutants such as PM2.5, PM10, NO2, SO2, CO, O3, Benzene, Toluene and Xylene
   b. crop_production_major_3_crops.csv – Agricultural Data (Unified Portal for Agricultural Statistics 2025) which contains the agricultural data such as crop, season and state
   c. Indian_festival_dates_2015_2020.xlsx – Festival Dates data contains festival dates of festivals such as Holi and Diwali from 2015 to 2020

2. Data Cleaning and Preprocessing: conversion of categorical variables to dummy variables, imputation of missing values for predictors using K-Nearest Neighbors (KNN) method due to its ability to leverage feature similarity, removal of colinear variables, merging of the datasets

3. Exploratory Data Analysis (EDA) and visualization:

   a. Determined the average AQI values and calculated the most polluted cities as Ahmedabad, Delhi and Patna with their respective AQI values 453.54, 260.15 and 237.92 and identified the months of October, November, December, January and February as the months with high AQI compared to other months as seen in *Figure 1*.

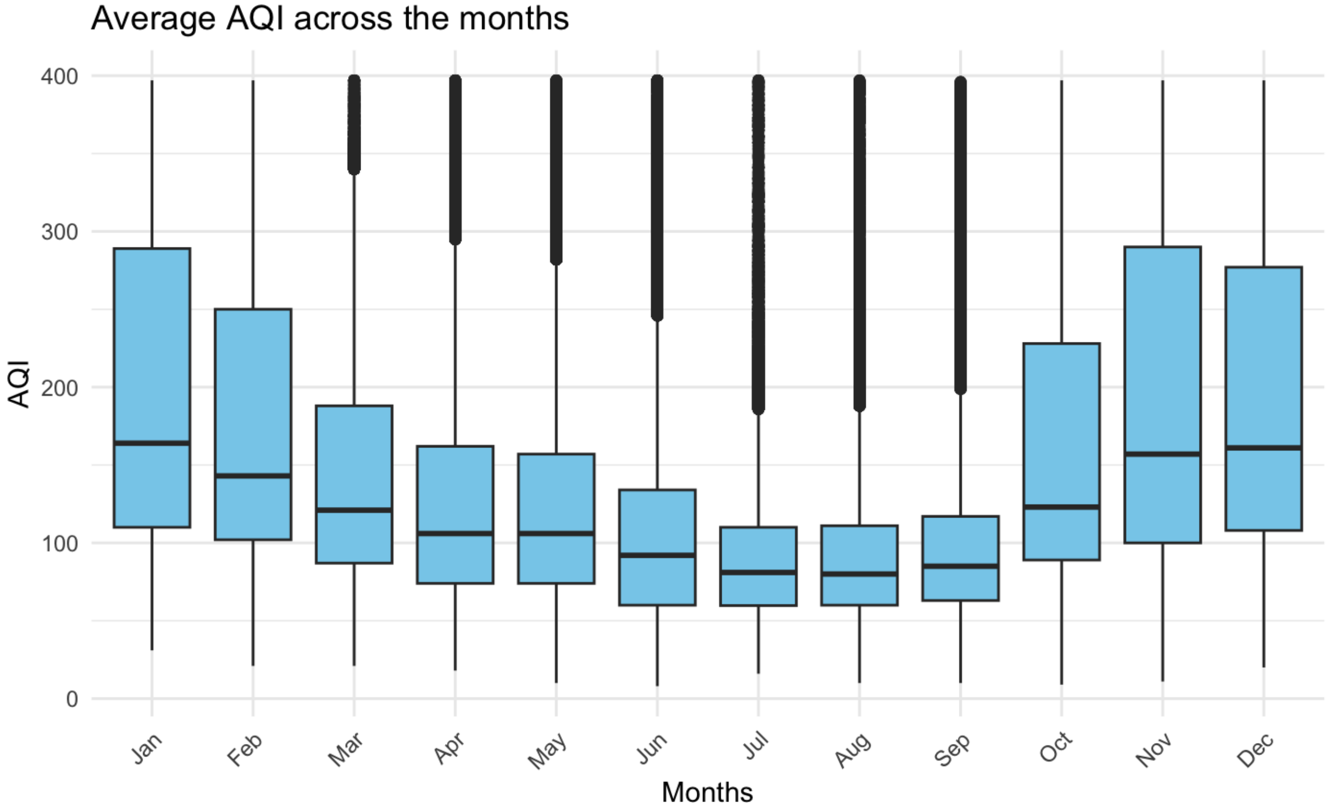

Figure 1: Boxplot Distribution of average AQI across the months from 2015 to 2020

   b. computed the correlation plot to determine if there a presence of multicollinearity between the predictors, and hence determined NOx, is highly like NO2 and NO (68%, 88%) as seen in *Figure 2 and 3*.

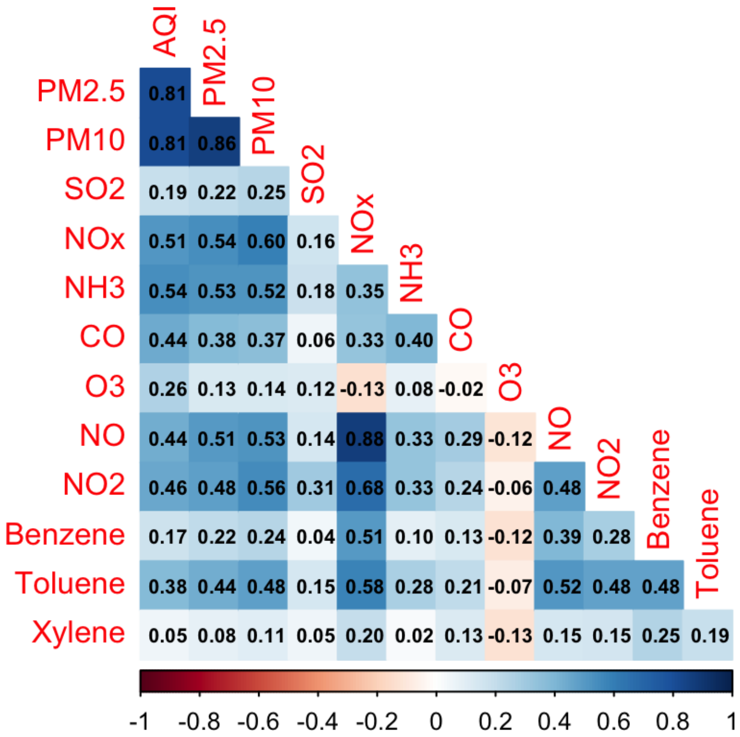

Figure 2: Correlation plot showing the relationships between pollutant variables used to predict AQI

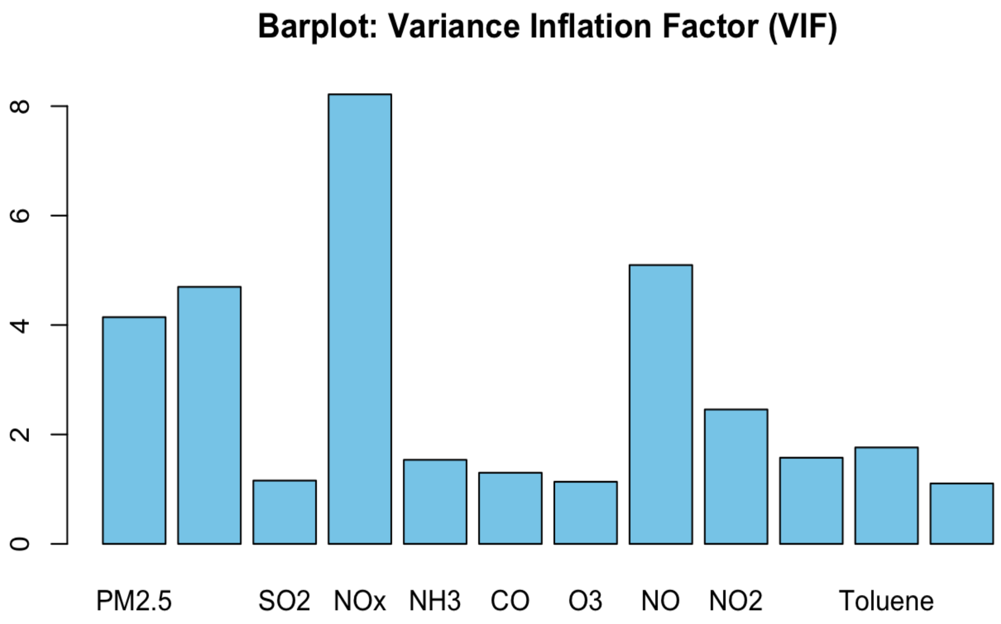

Figure 3: Barplot of Variance Inflation Factor (VIF) values indicating multicollinearity among pollutant predictors in the AQI regression model

c. variable importance plot helped determine that PM 10 and PM 2.5 are the most important features and NO2, Benzene, Xylene and Toluene are not important predictors in calculating AQI as seen in *Figure 4*.

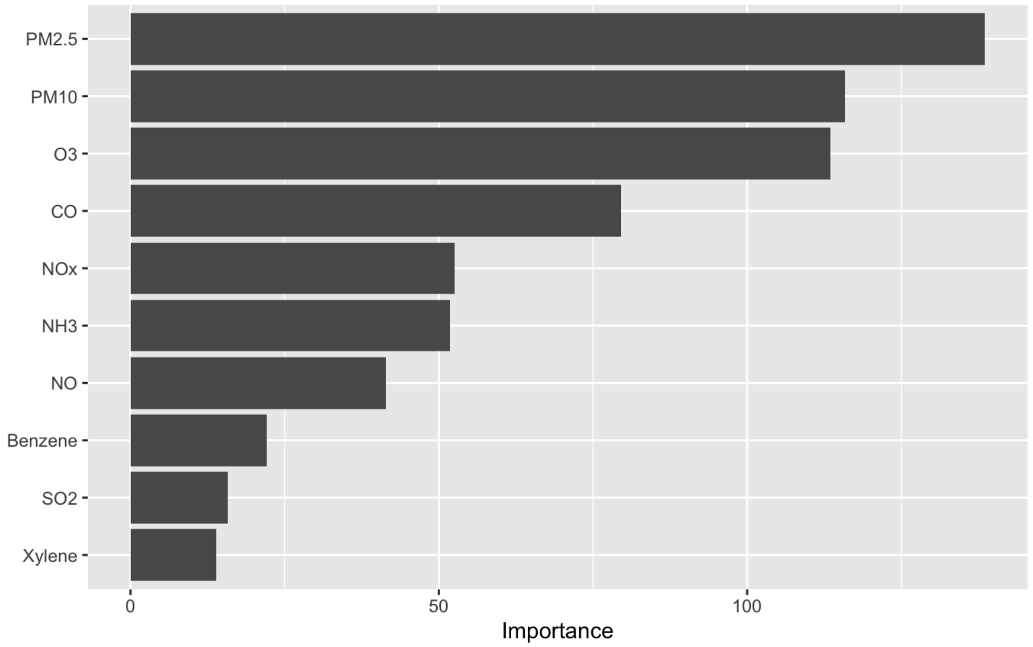

Figure 4: Variable Importance Plot depicting order of importance for predictor variables (Pollutants) on the response (AQI) in a linear regression model

4. ML Model Training
   To predict the AQI of top polluted cities in India, I have used the following models
   a. Linear Regression: It served as a baseline model to establish linear relationship between the pollutants such as PM 2.5, PM 10, Nox
      - Limitations: Unable to capture nonlinear relationships in the data
   b. Lasso Regression (L1 Regularization): helps in feature selection by shrinking the coefficients of irrelevant features to zero
      - Limitation: similar to linear regression, unable to capture nonlinear trends in the dataset
   c. Random Forest: it uses numerous decision trees and finally evaluates their combined outputs to make more accurate predictions than an individual decision tree.
   d. XGBoost (eXtreme Gradient Boosting): It constructs decision trees sequentially, with each tree rectifying faults made by the previously trained trees. (GeeksforGeeks, 2024)

5. ML Model Optimization
   Performing hyper parameter optimization using tuning grid and using cross validation so that the model can generalize on new data.

Model	Hyperparameter	Tuning Grid Values	Tuned Values
Lasso Regression	lambda	10^{-4} to 10^{4}	1.000230
Random forest	Mtry	1 to 10	10
	min_n	2 to 10	2
XGBoost	tree_depth	3 to 10	10
	learn_rate	0.001 and 0.1	1.002305

Table 2: Hyperparameters tuning grid and tuned values

6. ML Model Selection: after optimizing the ML models and evaluating their performance on both the training as well as testing datasets, XGBoost is selected with the hyperparameters tree depth 10, learn_rate 1.002305.

Model	Train RMSE	Test RMSE
XGBoost	20.7	121
Random Forest	48.2	120
Lasso Regression	248	249

Table 3: Performance of the different models on training and testing dataset

4. Discussion

Role of Events:

During the study it was established that the main predictors of AQI are PM 10 and PM 2.5, which are the main emissions from burning firecrackers and stubble of crops such as wheat and rice, other predictor values such as NO2, CO, and O3 are released in the atmosphere mainly due to burning fossil fuels and vehicular emissions who effects on AQI are beyond the scope of the report and can be considered in future study. Figure 5 clearly shows that there was an extreme increase in the AQI in October 2017 with Diwali on 19 October 2017 for Ahmedabad, however an opposite trend was seen in Delhi, this can be explained by firecracker and crop burning ban in and around Delhi which resulted in lower PM 2.5 and PM 10 values hence low AQI.

In Ahmedabad, similar high peaks can be observed in Nov 2018 (Diwali on 7 Nov), Oct 2019 (Diwali on 27 Oct) as compared to other festivals such as Dusshera, Lohri and New Year's Eve. Significant drop in the AQI values can be observed from October 2019, during the covid pandemic and both the states had their lowest value of AQI during this time due to restrictions of movements industry closure.

Policy Implications:

1. After the success of firecracker ban in 2017 the supreme court of India cleared that the ban was not only for Delhi and surrounding regions but for the entire country, however, it allowed for green cracker and those with reduced emissions (Vajiram Editor, 2018)
2. Increased campaign and awareness among farmers to deter them from stubble burning and imposing hefty fines for each incident of stubble burning under the new rules of Environment Protection Act (EPA)
3. Pushing for In-Situ Model instead of stubble burning taking the stubble of plants and mixing it in the soil with machines so that it can decompose and make the soil fertile instead of burning is a sustainable way of doing agriculture. Government is heavily subsidizing these machines, since 2018 government have distributed over 200,000 machines worth 2.4 billion.

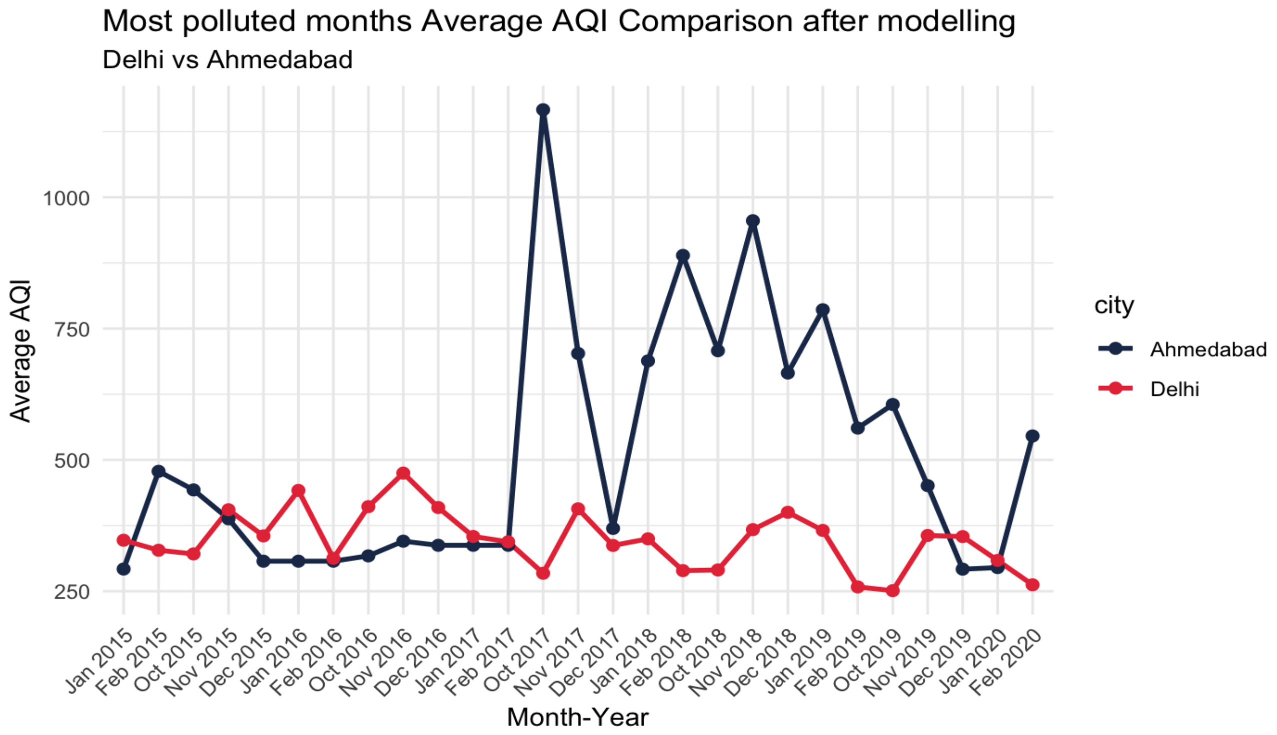

Figure 5: Monthly average AQI comparison between Delhi and Ahmedabad across high pollution months after modelling

5. Limitations

1. The report does not consider emissions from other sources such as vehicular emission, industrial emission, power plants and household fuel burning which also significantly impact the air quality
2. The dataset only covers the data from 2015 to 1 July 2020, limiting the analysis
3. The report is city specific and may not generalize to other cities of India due to geographic and climatic factors

6. Conclusion

The study was able to successfully identify PM 10 and PM 2.5 to be the main pollutants responsible for air pollution in Delhi and Ahmedabad. These particulate matters are the main components of emissions resulting from burning firecrackers and crop stubble. Comparing the predicted AQI trends in Delhi and Ahmedabad during the high pollution months revealed that Ahmedabad has severe peaks in AQI values during and around Diwali in the year 2017, 2018 and 2019. These spikes also coincided with rice stubble burning during the month of October and November which is the harvest period of 'Kharif rice' or winter rice. However, Delhi recorded lower pollution in the year 2017 on the accounts of firecracker and stubble burning bans set up by the government.

Among the festivals Diwali resulted in higher AQI values, followed by Dussehra. The Outbreak of Covid-19 pandemic resulted in a lockdown in both the cities resulting in restrictions on the movement of people (no festival celebration), vehicles and closed factories which in turn help reduce the air pollution. The government's policies and decision are also helping to curb the pollution by scaling up the firecracker ban to the entire country, levying hefty fines on stubble burning (Shagun 2024) and their decision to provide subsidized machines that can mix the crop stubble in the soil as an alternative to stubble burning. Overall, by correlating event driven variations with the pollutant data can help the policymakers better understand what works and design more effective measure to curb the air pollution.

7. Future Work

1. Updating the city_hour.csv dataset with updated values of pollutants after July 2020, particularly during March 2021 when the lockdown was lifted
2. Expanding the study of most polluted cities to other cities such as Patna, Kanpur
3. Supplementing the dataset with vehicular and industrial emission data to obtain a more holistic AQI predictions
4. Analyzing the long-term impact of policies of government on the air quality

8. Replication Package

https://github.com/nishantsingh98/Big-Data-Project

9. References

1. Rao, R 2020, _Air Quality Data in India (2015 - 2020)_, Kaggle, viewed 3 June 2025, <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>

2. Kumar, A 2025, _Pollution in India_, Encyclopedia Britannica, viewed 5 June 2025, <https://www.britannica.com/topic/pollution-in-India.>

3. Unified Portal for Agricultural Statistics: UPAg 2025, _APY Cropwise Insights_, UPAG, viewed 5 June 2025, <https://upag.gov.in/.>

4. Central Pollution Control Board n.d., _About National Air Quality Index_, Central Pollution Control Board, viewed 3 June 2025, <https://cpcb.nic.in/displaypdf.php?id=bmF0aW9uYWwtYWlyLXF1YWxpdHktaW5kZXgvQWJvdXRfQVFJLnBkZg>

5. Basu, M 2019, ‘The Great Smog of Delhi’, _Lung India_, vol. 36(3), pp. 239, DOI:10.4103/lungindia.lungindia_363_18.

6. Regan, H 2024, _Air pollution in Asia worsened in 2023, with Bangladesh, Pakistan, and India among the worst hit_, CNN, viewed 5 June 2025, <https://edition.cnn.com/2024/03/18/climate/air-pollution-report-2023-asia-climate-intl-hnk.>

7. Dybwad, A 2022, 'What to Know About India’s Crop Burning Season', _PurpleAir_, 16 November, viewed 1 July 2025, <https://www2.purpleair.com/blogs/blog-home/what-to-know-about-india-s-crop-burning-season#:~:text=India's%20crop%20burning%20season%20usually,rural%20areas%20to%20metropolitan%20cities>

8. Pandey, P 2020, _Indian cities database_, Kaggle, viewed 1 July 2025, <https://www.kaggle.com/datasets/parulpandey/indian-cities-database>

9. Chanana I, Sharma A, Kumar P, Kumar L, Kulshreshtha S, Kumar S, Patel SKS, 2023, 'Combustion and Stubble Burning: A Major Concern for the Environment and Human Health', _Fire_ 2023, 6, 79, DOI: https://doi.org/10.3390/fire6020079

10. Gouder, C., & Montefort, S 2014, 'Potential impact of fireworks on respiratory health', _Lung India_, 31(4), 375–379, DOI: https://doi.org/10.4103/0970-2113.142124

11. GeeksforGeeks (2024), _Difference Between Random Forest and XGBoost_, GeeksforGeeks, <https://www.geeksforgeeks.org/machine-learning/difference-between-random-forest-vs-xgboost/.>
12. Ganguli, I , Nakum, M., Das, B. & Kshetrimayum, N. 2025, *Comprehensive analysis of air quality trends in India using machine learning and deep learning models*, pp. 313–318, https://doi.org/10.1145/3700838.3703681.
13. Mujtaba, M.A., Munir, M.A., Ali, S., Petrů, J., Ansar, T., Akhlaq, W., Ahmad, M., Iqbal, H., Ali, F., Bashir, M.N. & Alexander, T. 2025, *Using machine learning for air quality prediction and sustainable urban planning*, *Sustainable Futures*, vol. 10, p. 100981, [online], https://doi.org/10.1016/j.sftr.2025.100981.
14. Pande, C.B., Kushwaha, N.L., Alawi, O.A., Sammen, L.S., Sidek, L.M., Yaseen, Z.M., Pal, S.C. & Katipoğlu, O.M. 2024, *Daily scale air quality index forecasting using bidirectional recurrent neural networks: Case study of Delhi, India*, *Environmental Pollution*, p. 124040, https://doi.org/10.1016/j.envpol.2024.124040.
15. Rahman, M.M., Nayeem, M.E.H., Ahmed, M.S., Tanha, K.A., Sakib, M.S.A., Uddin, K.M.M. & Babu, H.M.H. 2024, *AirNet: predictive machine learning model for air quality forecasting using web interface*, *Environmental Systems Research*, vol. 13, no. 1, https://doi.org/10.1186/s40068-024-00378-z.
16. Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A. & Sonne, C. 2023, *Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam*, *Chemosphere*, vol. 338, p. 139518, [online], https://doi.org/10.1016/j.chemosphere.2023.139518.
17. Vajiram Editor 2023, *SC says cracker order applies to country, not just NCR*, [online], Vajiram and Ravi, viewed 5 August 2025, https://vajiramandravi.com/current-affairs/sc-cracker/
18. World Health Organization (WHO) 2025, *Air quality, energy and health*, [online], World Health Organization, , viewed 5 August 2025, https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts
19. Shagun (2024). *Centre doubles fine for stubble burning; farmers to pay up to Rs 30,000*, Down To Earth, viewed 5 August 2025, https://www.downtoearth.org.in/air/centre-doubles-fine-for-stubble-burning-farmers-to-pay-up-to-rs-30000
20. ShareAmerica 2024, *India’s initiatives for cleaner air*, [online], ShareAmerica, viewed 5 August 2025, https://share.america.gov/indias-initiatives-for-cleaner-air/