India's Air

Analysis and Prediction of Air Quality in Indian cities

Tony P Joy (tonyjoy@iisc.ac.in), Ashutosh Mishra (ashutoshmish@iisc.ac.in), Vivek H N (vivekhn@iisc.ac.in) , Neeraj Shete (neerajshete@iisc.ac.in)

Table of Contents

Problem Background

Globally, air pollution is a silent killer. The air pollution levels in India are among the highest in the world, posing a heavy threat to the country's health and economy. All of India's 1.4 billion people are exposed to unhealthy levels of harmful pollutants - emanating from multiple sources. In this project we are trying to analyse and understand the levels of major air pollutants and the resultant AQI (Air Quality Index) across the major cities in India. Through this analysis we intend to identify the patterns and observations related to the AQI that can provide the data backing to prepare a search/ warn system for the citizens on the air quality at a given date and time of the year and remind them about wearing their N95 masks.

While the depleting air quality is a growing concern in India it doesn't seem to get the required attention both from the responsible sector and the sufferer sector. We believe that the relevant data and live warnings – when incorporated with the daily digital tools like news feeds, tweets and maps can help to bring the desired emphasize and drive actions.

Project Objectives

Analyse and Visualize:

- Distribution of pollutants across different cities
- The major pollutants for each city and compare the average value of each pollutant across the cities.
- Compare the levels of pollutants across cities for the years 2019 – 2023.
- Study the levels of different pollutants for each city.
- Identify the seasonal change of pollutant values across cities and identify the months where a particular pollutant level reaches their peak.
- Impact of different environmental parameters on AQI value for a given city
- Identify the best and worst months w.r.t air quality for each city.
- Identify the cities having higher ratio of good AQI.
- Yearly trend for vehicular and industrial pollution to see if we have any cities showing positive of negative trend.

Build ML models to:

- Predict the AQI for a given city given the pollutant features.
- Forecast AQI for a given city on a given day.
- Evaluate the performance metrics of different regression models and select the best model for deployment.
- Fine tune the best regression model through hyperparameter tuning.
- Realtime predict AQI for a given city and display it.

Data sources and data models

Data is downloaded from "Central Control Room for Air Quality Management - All India". Link: CCR (cpcb.gov.in)

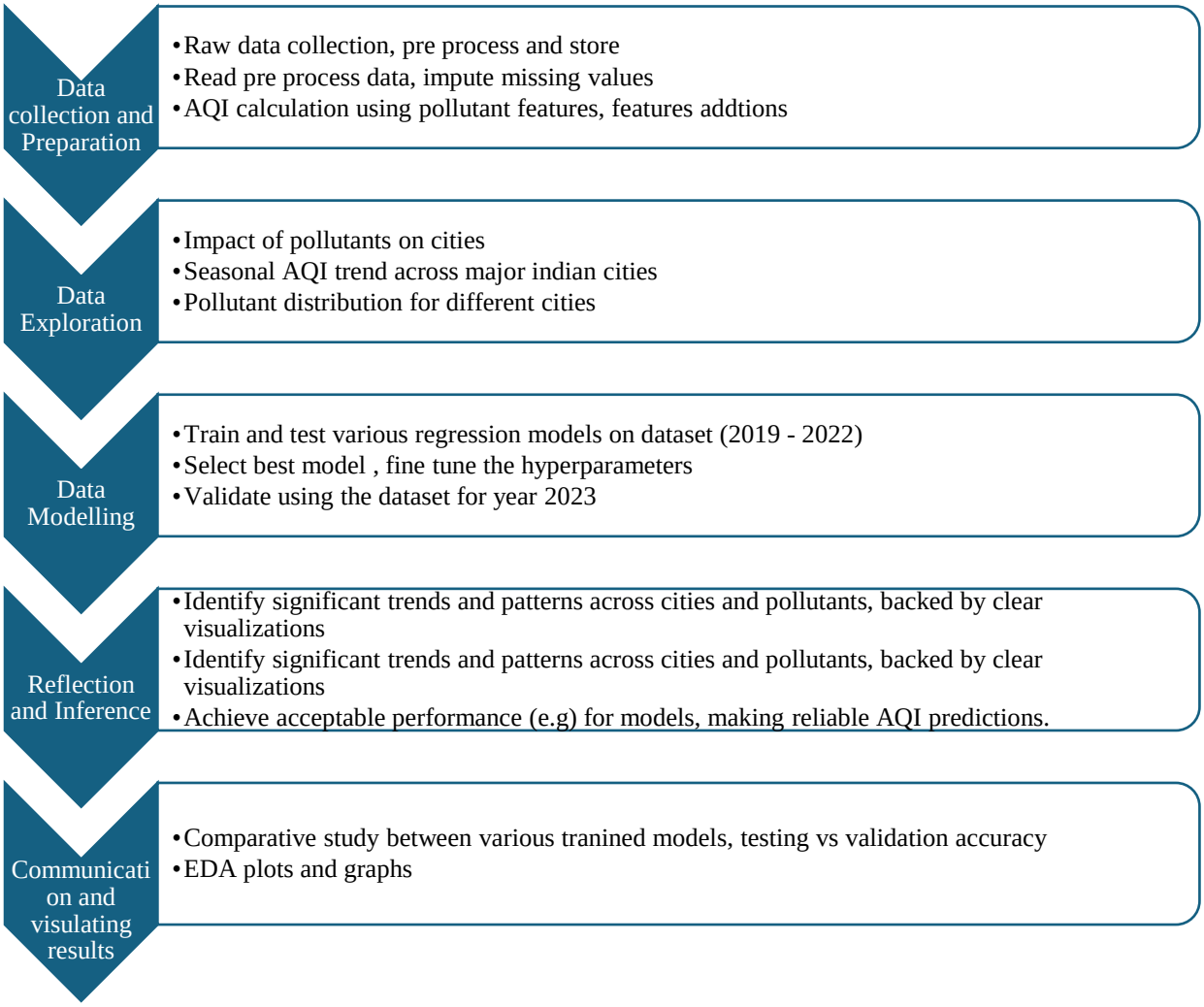
We are fetching data for 6 cities – Bengaluru, Hyderabad, Chennai, Mumbai, Kolkata, and Delhi. We are collecting yearly dataset for each of these cities for the years 2019 through 2023. The dataset contains air quality features collected on an hourly basis making it ~8k samples for a city per year.

Features available in the dataset:

Timestamp, Measurement of different pollutants *(PM2.5, PM10, NO, NO2, NOx, NH3, SO2, CO, Ozone, Benzene, Toluene, Xylene, O Xylene, Eth-Benzene, MP-Xylene), AT (°C), RH (%), WS (m/s), WD (deg), RF (mm), TOT-RF (mm), SR (W/mt2), BP (mmHg), VWS (m/s))*

There are ~40k samples per city and we are considering 6 cities.

High-level design



Data collection and preparation

Data Collection

The dataset for this study was obtained from the "Central Control Room for Air Quality Management - All India," hosted on the CPCB (CCR (cpcb.gov.in)) platform. This open-access government repository is well-maintained, providing a comprehensive collection of historical air quality data, making it ideal for our analysis. Six major Indian cities—Bengaluru, Hyderabad, Chennai, Mumbai, Kolkata, and Delhi—were selected for

focused study. The data spans the years 2019 through 2023, offering hourly records, resulting in approximately 8,000 samples per city per year.

Data Retrieval and Storage

Datasets were segmented into **training data (2019–2022)** and **testing data (2023)**. These yearly datasets were consolidated into master datasets for training and testing, stored on a shared **Google Drive** for easy access and collaboration.

Data Cleaning

- Threshold-Based Clean-up: Columns with more than 10% missing values were removed to ensure data quality and reduce noise in the dataset.
- Interpolation-Based Imputation: For the remaining missing values, linear interpolation was applied, leveraging the time series nature of the data for accurate gap filling. The timestamp column was also converted to datetime format for seamless temporal analysis.

Feature Engineering

Advanced processing steps were applied to extract valuable insights:

- **Rolling averages** were computed using sliding windows to smooth fluctuation.
  - For pollutant P (e.g., PM2.5, PM10, NO2, etc.) in a specific city:

$$P_{24hr\_avg}(t) = \frac{\sum_{i=t-23}^{t} P(i)}{n}$$

  - t: Current hour.
  - P(i): Pollutant value at hour i.
  - n: Number of non-missing pollutant values in the rolling window (must be >= min_periods =16).

- **Rolling maximums** were computed for pollutant P (e.g., CO, Ozone)

$$P_{8hr\_max}(t) = \max\{P(t-7), P(t-6), \ldots, P(t)\}$$

  - Rolling window size is 8 hours.
  - P(i): Pollutant value at hour i.
  - Minimum periods = 1 ensures at least one value is available in the window.

- **Sub-index calculations** for key pollutants contributed to enriched analysis.
  - The sub-index for a pollutant is calculated using a piecewise linear scaling based on concentration breakpoints, as defined by air quality standards. The general formula can be described as:

$$I_p = I_{low} + \left(\frac{C_p - C_{low}}{C_{high} - C_{low}} \right) \times (I_{high} - I_{low})$$

- Where:
  - Ip: Sub-index for the pollutant p.
  - Cp: Observed concentration of the pollutant p.
  - Clow: Lower concentration breakpoint corresponding to the observed range.
  - Chigh: Upper concentration breakpoint corresponding to the observed range.
  - Ilow : Sub-index corresponding to Clow.
  - Ihigh: Sub-index corresponding to Chigh.

- **AQI (Air Quality Index)** and **AQI classification** were derived for a holistic view of pollution levels.

Calculation of AQI

Date			Station	NSIT	
DD-MM-YYYY			City	Delhi	
			State	Delhi	
Pollutants		**concentration in µg/m3 (except for CO)**	**Sub-Index**		**Air Quality Index**
				check	
PM10	24-hr avg	121.00	114	1	
PM2.5	24-hr avg	34.00	57	1	
SO2	24-hr avg	0.00	0	0	
NO2	24-hr avg	8.00	10	1	AQI = 114
*CO (mg/m3)	max 8-hr	0.00	0	0	
O3	max 8-hr	57.00	57	1	
NH3	24-hr avg	34.00	9	1	

* Concentrations of minimum three pollutants are required; one of them should be PM10 or PM2.5
* The check displays "1" when a non-zero value is entered

Good (0–50)	Minimal Impact	Poor (201–300)	Breathing discomfort to people on prolonged exposure
Satisfactory (51–100)	Minor breathing discomfort to sensitive people	Very Poor (301–400)	Respiratory illness to the people on prolonged exposure
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults	Severe (>401)	Respiratory effects even on healthy people

- Contributions of **vehicular** and **industrial pollution** were estimated to understand sectoral impacts.
  - Vehicular pollution = PM2.5 + NOx + CO + SO2
  - Industrial Pollution = Ozone + Benzene + NH3 + SO2 + PM10

This meticulous data preparation laid the foundation for developing robust models to analyse and predict air pollution trends in Indian cities.

Data Exploration

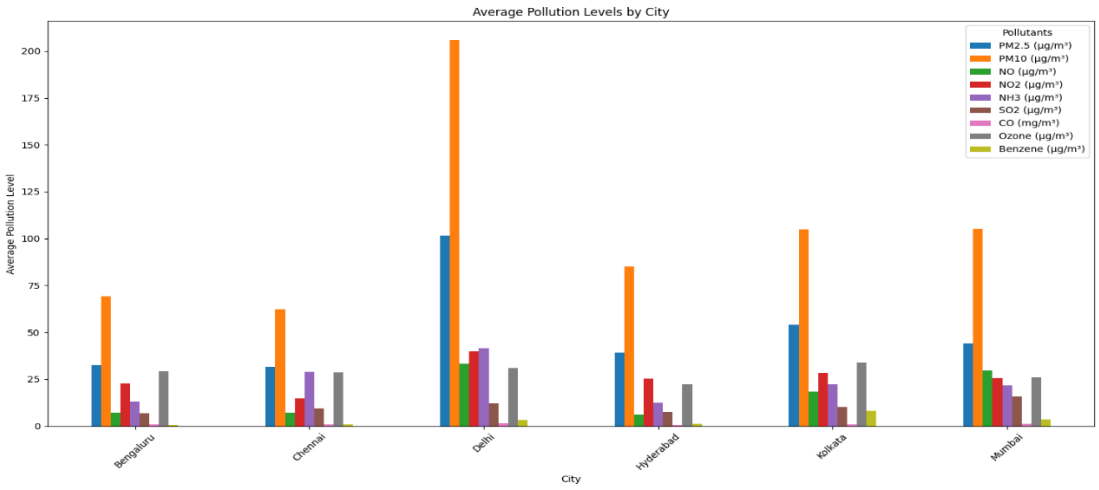
Impact on Environmental Parameters and AQI

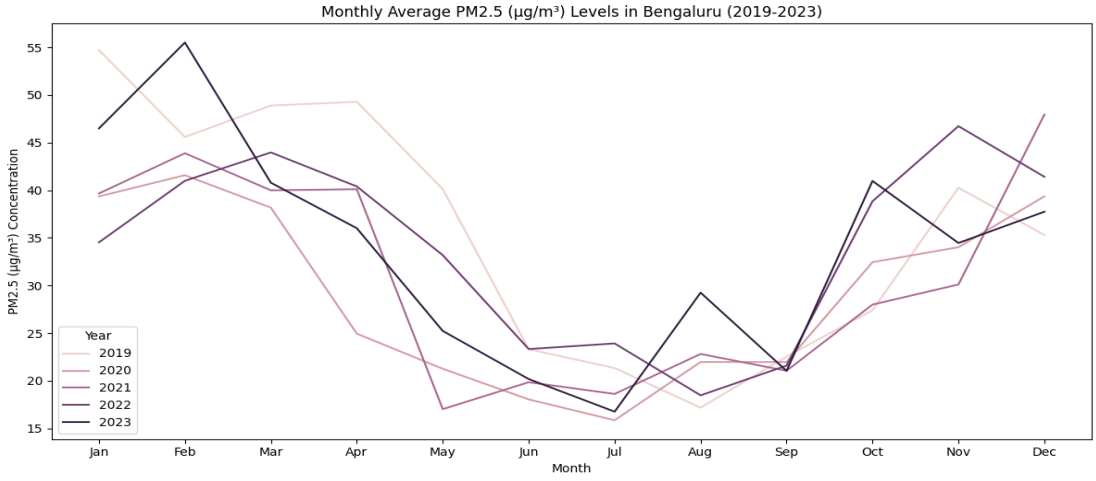
A detailed **correlation analysis** was conducted to understand the relationship between various environmental parameters and the Air Quality Index (AQI). This analysis helped identify key contributors to air pollution. The distribution of AQI classes was also examined, providing insights into the prevalence of different air quality categories across the dataset.

Distribution and Trends of Pollutants

City-wise plots were generated to visualize the distribution of pollutants, revealing the dominant pollutants in each city. Further analysis included:

- **Average levels** of major pollutants across cities.
- Year-over-year changes in pollutant levels to identify trends over time.
- **Monthly averages** and **seasonal variations**, highlighting peak pollution periods for each city.

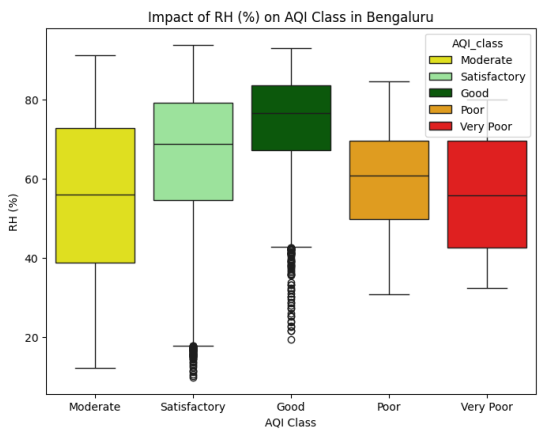


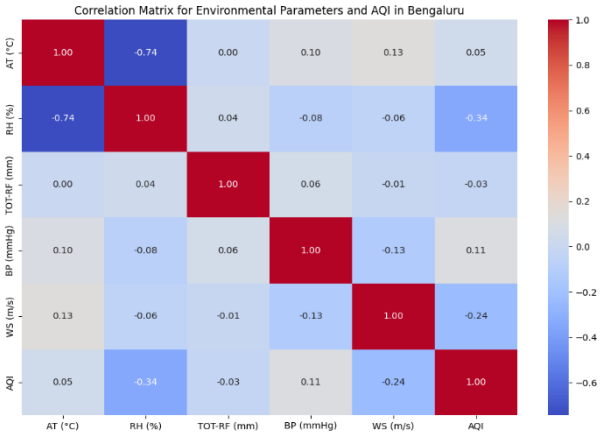


Environmental Parameters and City-specific Trends

Environmental parameters were analysed on a city-wise basis, focusing on:

- **Monthly averages** to identify recurring patterns.
- **Seasonal trends** to understand how environmental factors vary with seasons and their impact on air quality.

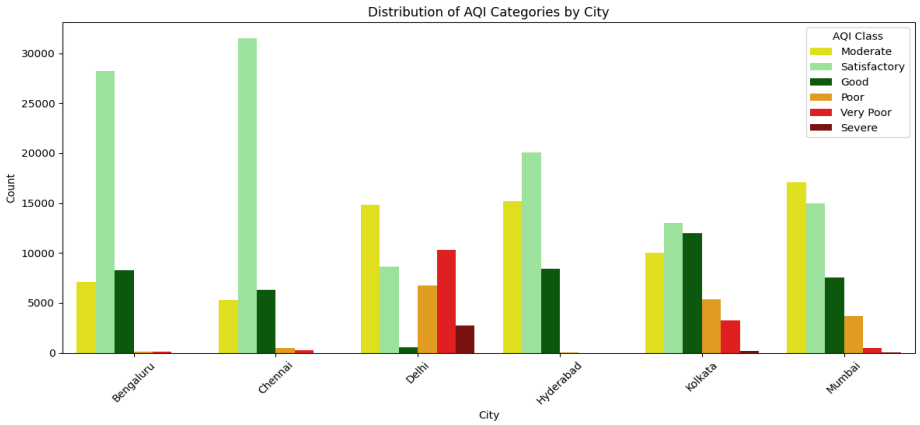


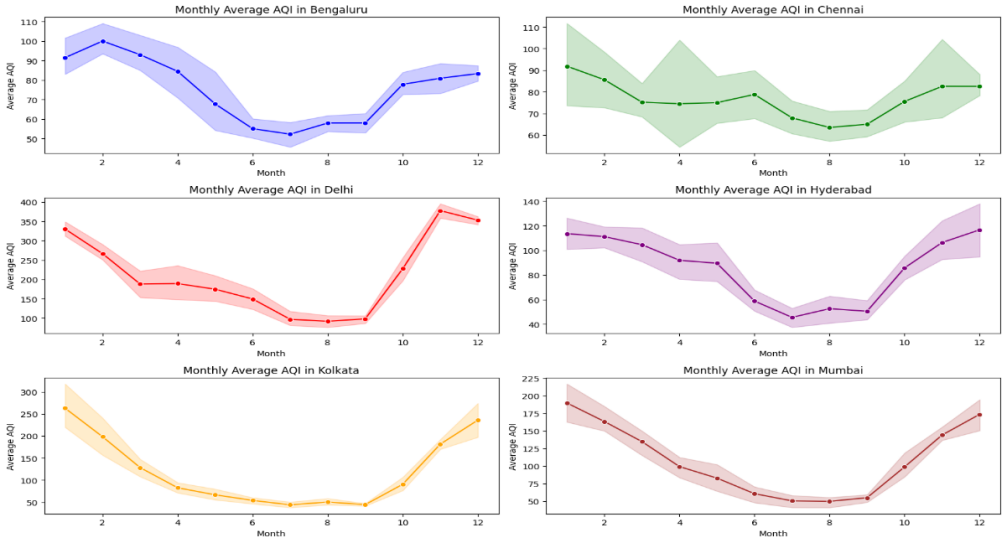


Analysis of City-wise AQI Values

City-specific AQI values were explored through:

- **Monthly AQI distributions**, showcasing variability throughout the year.
- **City-wise AQI comparisons**, identifying cities with the highest and lowest air quality levels.

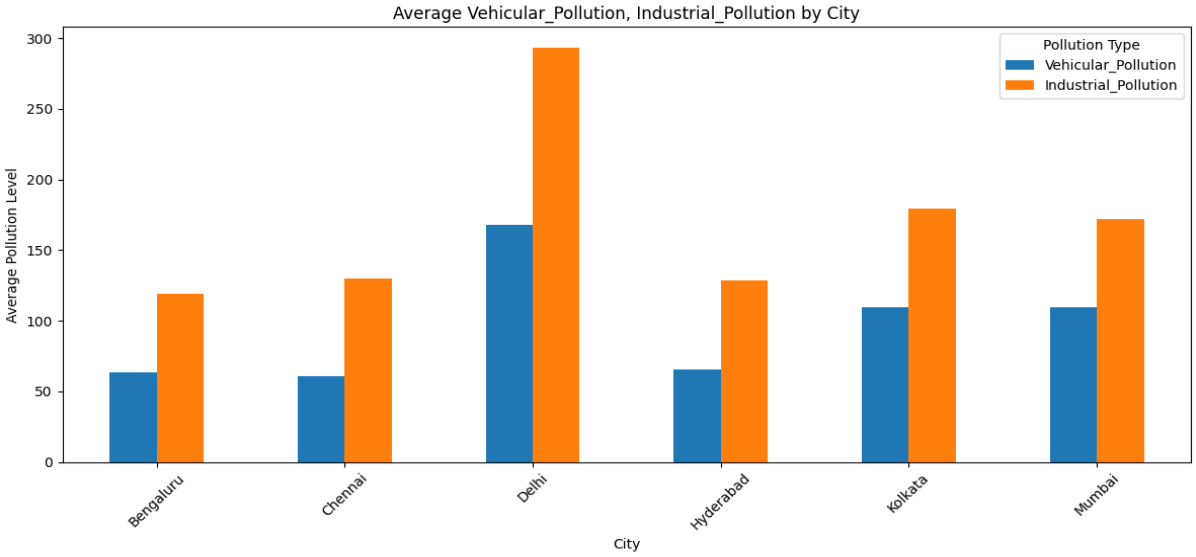




Vehicular and Industrial Pollution Studies

To understand the contribution of anthropogenic activities:

- The distribution of **vehicular and industrial pollution** was analysed for each city.
- **Monthly and yearly average trends** were studied, highlighting how these sources vary over time and their influence on air pollution patterns.



This exploratory analysis provided valuable insights into spatial, temporal, and seasonal air quality dynamics, forming a basis for deeper analysis and predictive modelling.

Data Modelling

The modelling process focused on developing, training, and comparing a variety of machine learning models to predict air quality trends. The dataset, spanning **2019–2022**, was used for training and testing, while the **2023 dataset** served as a validation set.

Model Selection

Four categories of models were explored:

1. **Linear Regression** techniques, including Lasso and Ridge, to capture linear relationships.
2. **Tree-based Regression** models like Decision Trees and Random Forests for their ability to handle non-linear patterns.
3. **Time Series Forecasting** using PyTorch's LSTM to model temporal dependencies in the data.
4. **Ensemble Models**, such as Random Forest (with Grid Search and Randomized Search for hyperparameter tuning), CatBoost, and XGBoost, known for their robustness and high predictive power.

Evaluation Metrics

Each model was assessed using **Root Mean Squared Error (RMSE)** and **training/testing accuracy** to measure performance. After training and testing, the best-performing model was selected and subjected to hyperparameter tuning to optimize its performance.

Validation

The selected best model (Random Forest regressor) was validated on the **2023 dataset**, with its accuracy on unseen data providing a final measure of its predictive reliability. This structured approach ensured comprehensive comparison and fine-tuning of models, enabling the identification of the most suitable algorithm for forecasting air quality. The process emphasized iterative improvement, leveraging both simple and advanced techniques for optimal results.

Reflection and Inference:

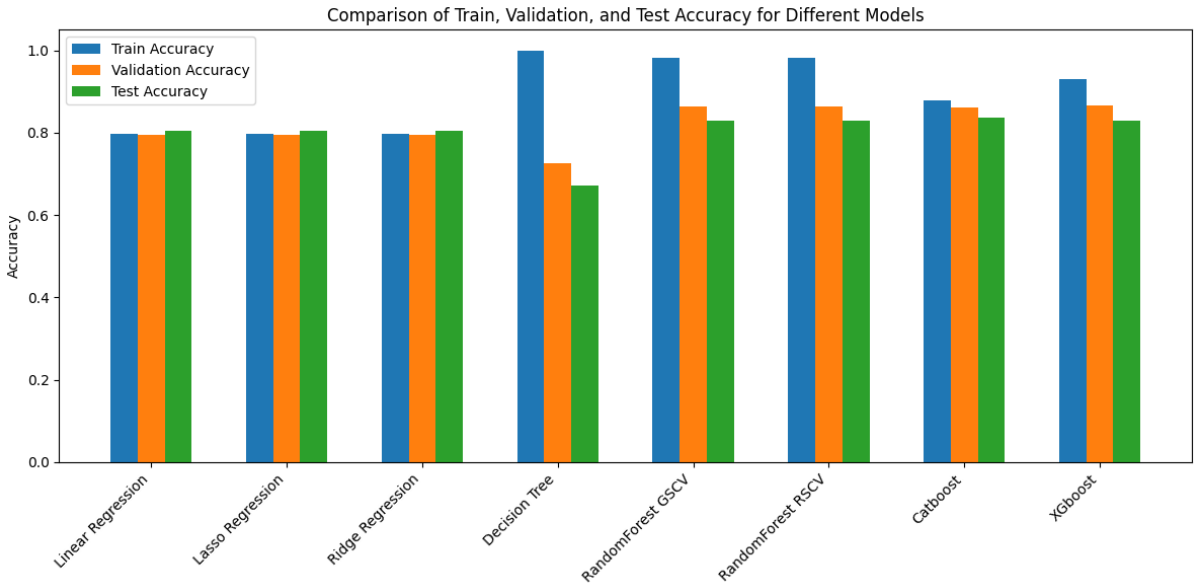
EDA and data Preparation metrics:

Data Preparation and EDA metrics	Success Criteria	Results Shown
Data Cleaning and Preprocessing	Achieve 90%+ clean, usable data across selected cities and pollutants.	Successfully cleaned the dataset with 0% missing values.
Distribution of Pollutants Across Cities	Understand the distribution of major pollutants for comparison across cities.	Observed flatter distributions in Delhi for PM pollutants, lower values in Bengaluru and Chennai.
Average Value of Major Pollutants for Each City	Visualize and compare average pollutant values across cities.	Found airborne PM as the dominant pollutant; Delhi and Mumbai had the highest averages for most pollutants.
Impact of Environmental Parameters on AQI	Identify relationships between AQI and environmental parameters like temperature, humidity, and wind.	Poor AQI linked to lower temperatures and higher humidity; higher wind speeds correlated with reduced pollution. .
Monthly Pollutant Levels (Seasonal Trends)	Identify seasonal changes and peak months for pollutants.	PM pollutants peaked in winter; NH3 and SO2 levels were highest during summer; monsoon reduced pollution.
AQI Trends Across Cities	Assess monthly AQI trends and identify best and worst months for air quality.	Delhi worsened in winter; June to September showed healthy AQI in most cities except Delhi.
Industrial vs. Vehicular Pollution	Determine dominant pollution sources for each city and their trends.	Industrial pollution exceeded vehicular in all cities; Delhi showed consistently high vehicular pollution levels.
City-Specific Insights	Highlight unique trends for each city.	Chennai and Bengaluru maintained better air quality; Delhi and Mumbai had higher pollutant levels and poor AQI.

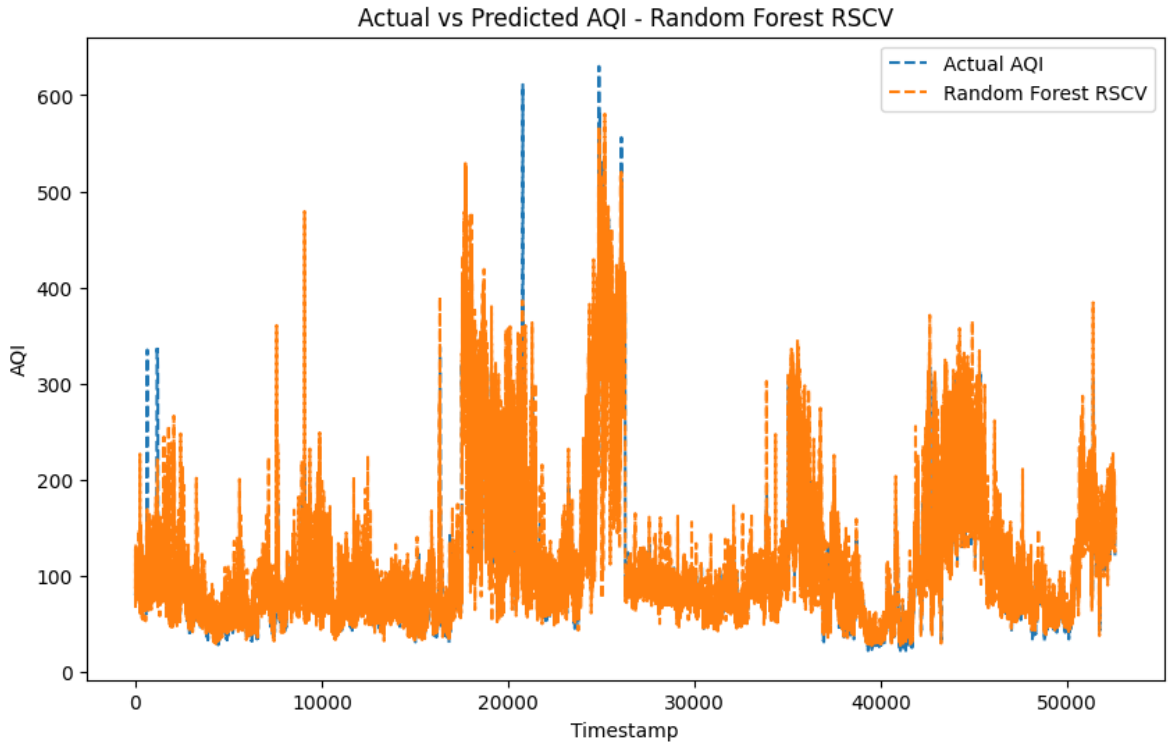
Pollutant Trends (2019-2024)	Compare pollutant trends over the years and identify long-term patterns.	Chennai showed sustained control; Mumbai exhibited upward trends in pollutants; Delhi remained consistently high.

Model Evaluation Metrics:

Models	RMSE		Training accuracy		Testing Accuracy		Validation accuracy	
	Achieved (%)	Target (%)	Achieved (%)	Target (%)	Achieved (%)	Target (%)	Achieved (%)	Target (%)
Linear Regression	39.07	35	79.8	80	80.5	75	79.5	75
Lasso Regression	39.08	35	79.8	80	80.6	75	79.5	75
Ridge Regression	39.06	35	79.8	90	80.5	75	79.5	75
Decision tree regressor	44.94	35	100	90	67.3	70	72.6	70
Random forest with Grid search	31.48	35	98.2	90	83	80	86.4	85
Random Forest with randomized search	31.38	35	98.2	90	83	80	86.4	85
CatBoost	32.1941	35	87.9	90	83.6	80	86.2	85
XgBoost	31.79	35	93	90	83	80	86.5	85

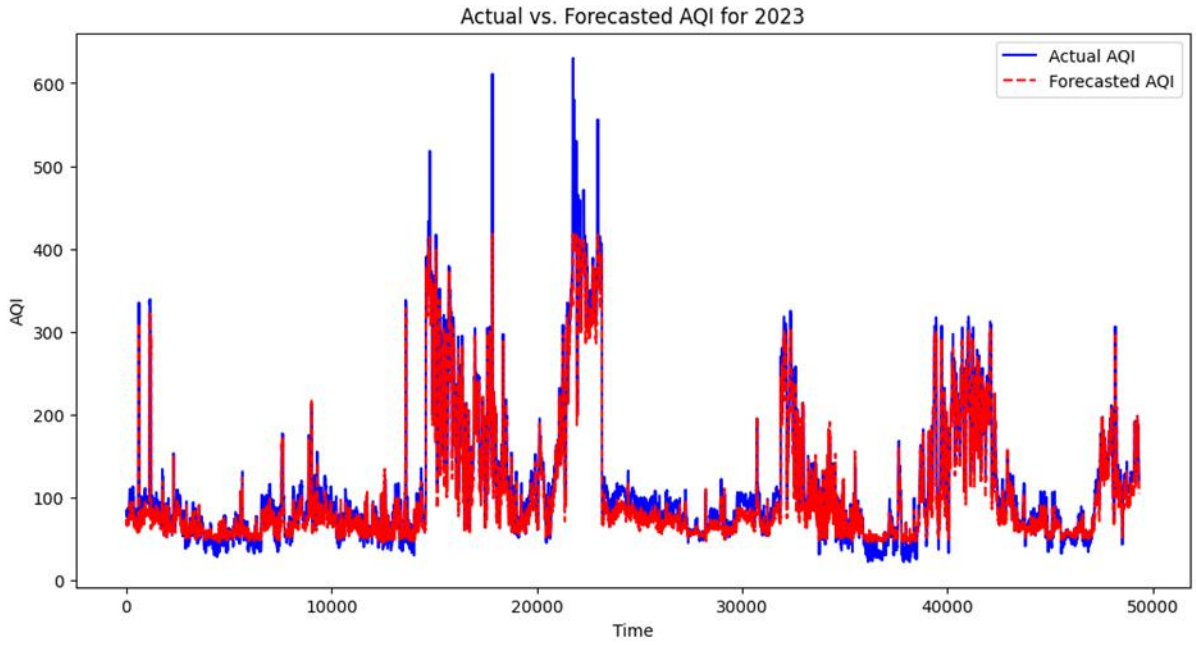


Best model's (Actual vs Predicted values):



LSTM Forecasting Accuracy:

Forecasting hourly AQI values for all samples in 2023 test dataset using PyTorch LSTM Model.

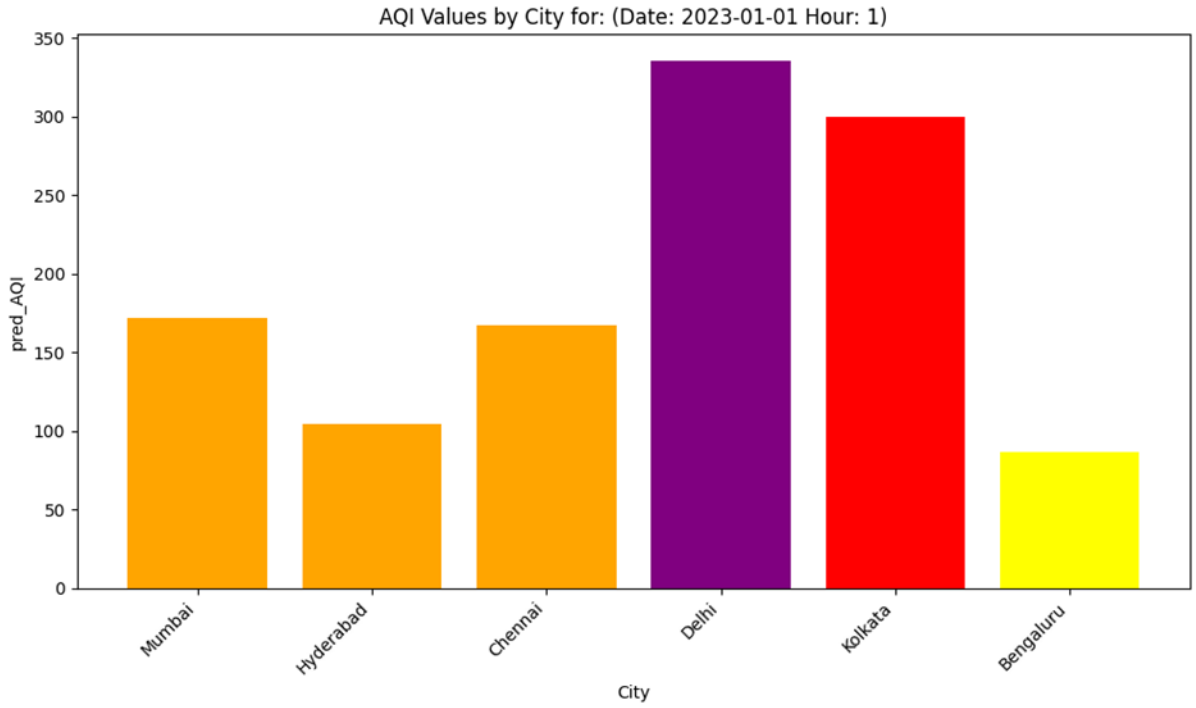


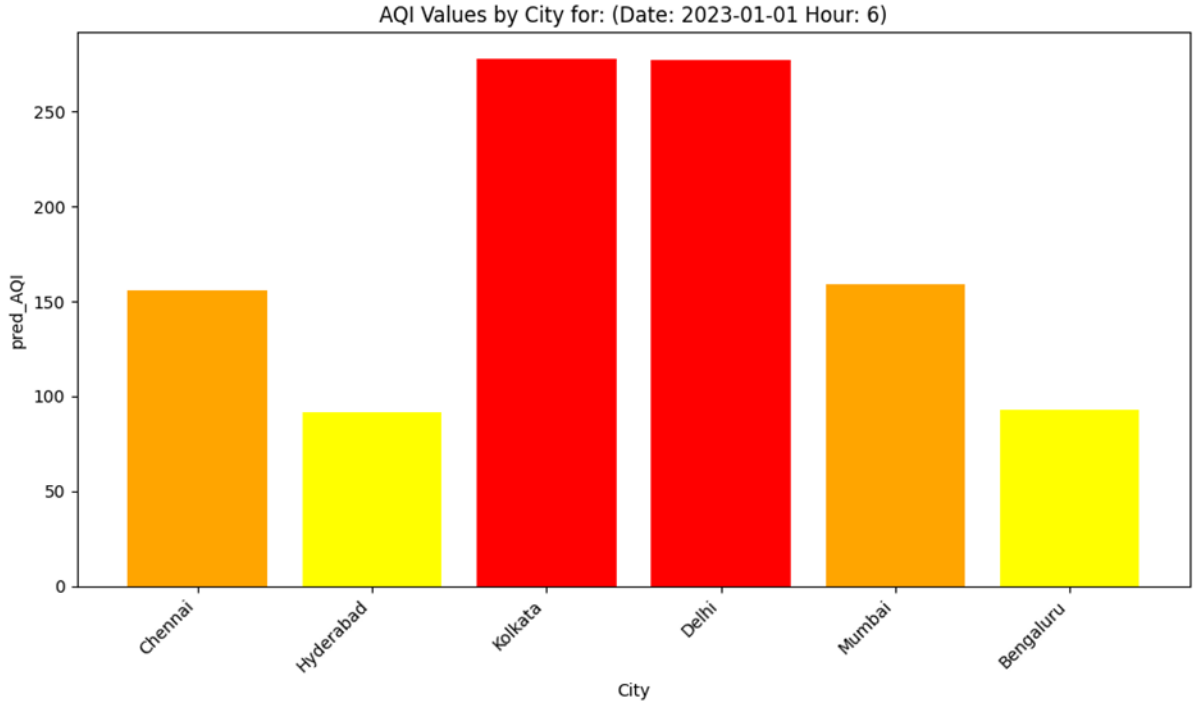
Deployment for Demo

The deployment of the AQI prediction model is designed for real-time monitoring and processing of new data. The key components and workflow are as follows:

1. **Input Data Source**: A shared directory serves as the input store where new CSV files containing hourly pollutant measurements for selected Indian cities are added. Each file represents a batch of new data to be processed. This is simulated using a file streaming thread that will stream a new CSV file to the input store from the 2023 test dataset grouped to hour of days effectively simulating hourly updates coming from the edge.
2. **Application Workflow**:
    a. A monitoring application continuously watches the input directory for new files.
    b. Upon detecting a new file, the application reads the data, validates it, and preprocesses it to match the model's input requirements.
3. **Prediction Process**:
    a. The application uses the saved best-performing model to predict the AQI value for each city in the new data.
    b. Predictions include the AQI value along with the associated date and hour from the input data.
4. **Visualization**:
    a. A bar plot is generated to display the predicted AQI values for all cities included in the batch. The plot highlights the date and hour of the predictions for clear temporal context.

This deployment setup ensures a streamlined and automated prediction pipeline, enhancing the usability and efficiency of the model in real-world applications.





Summary and Future enhancements:

The report outlines a project focused on analyzing and predicting air quality in six major Indian cities (Delhi, Mumbai, Bengaluru, Hyderabad, Chennai, and Kolkata) using hourly data from 2019 to 2023. It aims to understand pollutant distributions, seasonal trends, and their impact on AQI. The project employs machine learning models like Linear regression, Lasso and Ridge regression, Decision Tree regression, Random Forest regression, CatBoost and XGBoost regression for AQI prediction and forecasting.

Key deliverables include pollutant analysis, AQI trend visualization, and identifying seasonal and city-specific pollution patterns. Future enhancements propose developing a live dashboard for real-time AQI predictions and integrating alerts with digital platforms to raise awareness and drive actionable insights.

The document envisions significant future enhancements, including the development of a live dashboard to monitor and predict air quality in real time. This dashboard would integrate historical data with live feeds from air quality monitoring stations, offering dynamic visualization of pollutant levels and AQI trends across Indian cities. It aims to provide actionable insights through features like:

- Real-Time AQI Display: A constantly updated AQI index for each city, allowing users to view current pollution levels and compare them across locations.
- Predictive Analytics: Integration of machine learning models to forecast AQI for upcoming hours or days, providing early warnings for poor air quality.
- Personalized Alerts: Location-based notifications to alert users about high pollution levels in their vicinity, paired with protective recommendations such as mask usage or reduced outdoor activities.