SRI VENKATESWARA INTERNSHIP PROGRAM FOR RESEARCH IN ACADEMICS (SRI-VIPRA)

SRI-VIPRA

Project Report of 2025: SVP-2523

"Transforming Air Pollution Management in India with AI and ML Technologies."

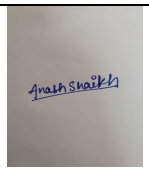
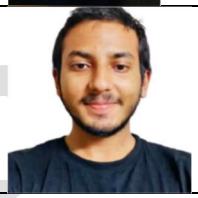
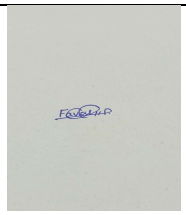
IQAC
Sri Venkateswara College
University of Delhi
Benito Juarez Road, Dhaula Kuan, New Delhi
New Delhi -110021

SRIVIPRA PROJECT 2025

Title : …"Transforming Air Pollution Management in India with AI and ML Technologies." **……..**

Name of Mentor: Dr. Tarakeswara Rao K **Name of Department:** Electronics **Designation:** Assistant Professor	

List of students under the SRIVIPRA Project

S.No	Photo	Name of the student	Roll number	Course	Signature
1		Anash Shaikh		Biological Sciences (H)	
2		Jawad Alamgir Purna		B.Sc. (Hons) Physics	
3		Muhammed Favas AK		B.Sc. (Hons) Biochemistry	
4					

5					
6					

Signature of Mentor

Certificate of Originality

This is to certify that the aforementioned students from Sri Venkateswara College have participated in the summer project SVP-2523 titled "-Transforming Air Pollution Management in India with AI and ML Technologies**".** The participants have carried out the research project work under my guidance and supervision from 1st July, 2025 to 30th September 2025. The work carried out is original and carried out in an online/offline/hybrid mode.

Signature of Mentor

Acknowledgements

We would like to express our gratitude and thanks to our professor and mentor, Dr. Tarakeswara Rao Kaviti for guiding us in this project. They provided us with their valuable suggestions throughout the process. They encouraged us to undertake this project from time to time.

We would also like to give special thanks to Sri Venkateswara College for providing all necessary facilities to carry out this Project smoothly. I will keep on trusting the College facilities for my future endeavours in the college.

We would like to express our sincere thanks to the Principal of Sri Venkateswara College, University of Delhi for organizing the SRI VIPRA internship Program, 2025. This program has helped us to put our scientific and computational understanding to application.

We would also like to thank the coordinators of SRI VIPRA program 2025 for providing us the opportunity to be a part of this SRI VIPRA PROJECT 2025.

We would also like to thank our family and friends for supporting us throughout this project.

TABLE OF CONTENTS

Executive Summary of the project

Air pollution is a pervasive and critical public health and environmental crisis in India, with numerous cities consistently ranking among the most polluted globally.
Traditional air quality management, reliant on sparse monitoring networks and reactive measures, has proven inadequate to address the complex, multi-source, and dynamic nature of the problem.
This report proposes a paradigm shift by integrating Artificial Intelligence (AI) and Machine Learning (ML) technologies into India's air pollution management framework. We demonstrate that AI/ML can transform the approach from reactive to **predictive**, from broad to **precise**, and from delayed to **proactive**.
Key findings indicate that AI/ML models can achieve high-accuracy:

- **Short-term (72-hour) and seasonal forecasts** of air quality indices (AQI).
- **Source apportionment** in near real-time, identifying contributions from vehicles, industry, dust, and stubble burning.
- **Hyperlocal pollution hotspot** mapping at a 1km x 1km resolution or finer.

The report outlines a strategic implementation framework, including data acquisition, model development, and integration with policy mechanisms like the Graded Response Action Plan (GRAP). It also critically examines challenges such as data quality, computational infrastructure, and interdisciplinary collaboration. The conclusion asserts that the adoption of AI/ML is not merely an upgrade but a necessary evolution for India to achieve its clean air goals, safeguard public health, and ensure sustainable economic development.

---

1. Introduction

1.1. The Indian Air Pollution Crisis

India faces a severe air pollution challenge, with pollutants like PM2.5 and PM10, Nitrogen Oxides (NOx), and Ozone (O3) consistently exceeding national and international safety standards. The annual average PM2.5 concentration in many northern Indian cities is often 5-10 times the WHO guideline. This has dire consequences, including millions of premature deaths, increased respiratory and cardiovascular diseases, and significant economic losses.

1.2. Limitations of Current Management Systems

The existing framework, while improved with the National Clean Air Programme (NCAP), relies heavily on:

- **Sparse Physical Monitoring:** The number of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) is insufficient for a country of India's size, creating data gaps.
- **Reactive Measures:** Actions under GRAP are often triggered after pollution levels have already peaked.
- **Delayed and Approximate Source Analysis:** Traditional source apportionment studies are expensive, time-consuming (taking months), and provide a snapshot rather than a dynamic view.
- **One-Size-Fits-All Policies:** Broad regulatory measures may not be effective for hyperlocal pollution hotspots.

1.3. The Promise of AI and ML

AI and ML offer the computational power to analyse vast, heterogeneous datasets and uncover complex, non-linear patterns that are impossible for traditional models to detect. They can learn from historical and real-time data to provide actionable insights, transforming air quality management into a data-driven science.

2. Literature Review

Studies have shown the efficacy of ML models like **Long Short-Term Memory (LSTM)** networks and **Random Forests** in forecasting PM2.5 levels with over 85% accuracy. Globally, China's use of AI for its "war on pollution" and IBM's Green Horizon project are notable examples. In India, research from IITs and start-ups has begun demonstrating the value of ML in predicting Delhi's winter pollution, primarily linking it to meteorological data and stubble burning fire counts.

2.1. Introduction

Air pollution poses a catastrophic public health and environmental crisis in India. Numerous Indian cities consistently rank among the most polluted globally, with pollutants like PM2.5, PM10, NO2, SO2, and O3 exceeding safe limits by a significant margin (WHO, 2021). The conventional approach to air quality management has relied on sparse monitoring networks, source apportionment studies, and regulatory frameworks, which, while valuable, are often reactive, slow, and lack the granularity and predictive power required for effective, proactive intervention. This review synthesizes existing literature on air pollution in India, the limitations of current management strategies, and the burgeoning potential of Artificial Intelligence (AI) and Machine Learning (ML) technologies to revolutionize this domain.

2.2. The Indian Air Pollution Context: Sources, Dynamics, and Current Monitoring

2.2.1 Pollution Sources and Seasonal Variation

The literature unequivocally identifies the primary sources of air pollution in India as a complex mix of:

- Anthropogenic Sources: Vehicular emissions, industrial discharge, construction dust, power generation (especially from coal), and the widespread practice of agricultural residue burning (Guttikunda & Calori, 2017).
- Natural Sources: Dust storms, particularly in the northern Indo-Gangetic Plain. A critical characteristic is the strong seasonal variation. The post-monsoon period (October-November) sees a dramatic spike in PM2.5 levels due to agricultural

burning (stubble burning) in states like Punjab and Haryana, compounded by unfavourable meteorological conditions (low wind speed, low temperature, and inversion layers) that trap pollutants (Jethva et al., 2018).

2.2.2 Limitations of Current Monitoring and Management

India's primary monitoring network, the Central Pollution Control Board's (CPCB) National Air Quality Monitoring Programme (NAMP), while expanding, faces several challenges:

- Spatial Sparsity: Fixed monitoring stations are expensive to install and maintain, leading to a sparse network that cannot capture hyper-local variations in pollution (Kumar et al., 2020).
- Data Latency: Data is often reported with a lag, making real-time public alerts and immediate policy actions difficult.
- Reactive Policy: Policies like the Graded Response Action Plan (GRAP) are often triggered based on current air quality indices, making them reactive rather than proactive.
- Complex Source Apportionment: Traditional source apportionment studies are resource-intensive, time-consuming, and provide a snapshot in time, failing to capture dynamic changes in source contributions.

2.3. The Emergence of AI and ML in Environmental Science

AI and ML have emerged as powerful tools for handling complex, non-linear, and high-dimensional data, making them exceptionally suited for environmental modelling.

2.3.1 Foundational ML Techniques for Air Quality

- Predictive Modeling: Regression algorithms (Linear, Lasso, Ridge), Decision Trees, and ensemble methods like Random Forest and Gradient Boosting Machines (e.g., XGBoost) have been widely used to predict pollutant concentrations (AQI) based on historical air quality data and meteorological parameters (temperature, humidity, wind speed/direction, etc.) (Masih, 2019).
- Time-Series Forecasting: Models like ARIMA (Auto-Regressive Integrated Moving Average) and, more recently, LSTMs (Long Short-Term Memory

networks) have shown superior performance in capturing temporal dependencies for forecasting pollutant levels hours or days in advance (Bai et al., 2020).

2.3.2 Advanced AI Applications for Pollution Management

Beyond forecasting, the literature points to more sophisticated applications:

- Spatio-Temporal Analysis: Geospatial AI (GeoAI) techniques, combined with satellite data (e.g., from MODIS, Sentinel-5P), can fill the gaps left by sparse ground monitors. Models can create high-resolution pollution concentration maps, identifying pollution hotspots at a city-block level (Shtein et al., 2020).
- Source Identification and Apportionment: ML algorithms can analyze the chemical composition of PM2.5 and correlate it with known source profiles to continuously estimate the contribution from different sources (vehicles, dust, biomass burning, etc.) (Sharma et al., 2022).
- Hypothetical Scenario Analysis ("What-If" Scenarios): AI systems can model the potential impact of various intervention strategies, such as the effect of banning diesel vehicles in a specific area or reducing industrial output by a certain percentage, providing data-driven insights for policymakers.

2.4. Review of AI/ML Applications in the Indian Context: Progress and Gaps

A growing body of research applies AI/ML to Indian air quality data, demonstrating promising results but also revealing significant gaps.

2.4.1 Evidence of Success

- Forecasting: Studies in Delhi, Chennai, and other metropolitan areas have successfully used LSTM and XGBoost models to achieve high-accuracy short-term (24-72 hour) AQI forecasts, outperforming traditional statistical models (Kumar & Pande, 2022).
- Source Analysis: Research using Positive Matrix Factorization (PMF) combined with ML techniques has provided more dynamic insights into the changing contributions of stubble burning to Delhi's winter pollution (Singh et al., 2021).
- Data Fusion: Projects have integrated data from low-cost sensors (LCS) with satellite imagery and meteorological data using ML to create more detailed air

quality maps, demonstrating the feasibility of a denser, virtual monitoring network.

2.4.2 Identified Research Gaps

Despite this progress, the literature reveals critical gaps that this project aims to address:

1. Lack of Integrated, Real-Time Decision Support Systems (DSS): Most studies remain academic exercises. There is a scarcity of operational, integrated AI platforms that provide real-time forecasts, source attribution, and policy recommendations in a single dashboard for urban local bodies and regulators.
2. Insufficient Focus on Source-Specific Interventions: While models can predict AQI, there is a need for systems that can directly link forecasts to actionable, source-specific mitigation strategies (e.g., predicting stubble burning plume trajectories and recommending targeted farm-level interventions).
3. Limited Exploration of Transfer Learning and Generalizability: Models are often city-specific. The potential of transfer learning to create robust models that can be adapted quickly to new Indian cities with limited data is underexplored.
4. Underutilization of Multimodal Data: The integration of novel data sources like traffic camera feeds, social media sentiment, and industrial energy consumption data with traditional environmental data for refined modeling is still in its nascent stages in India.

2.5. Conclusion and Project Positioning

The literature confirms that air pollution in India is a severe, multi-faceted problem that existing management paradigms are struggling to contain. Concurrently, AI and ML technologies have matured and demonstrated their potent capabilities in forecasting, pattern recognition, and data fusion within environmental science.

This project, "Transforming Air Pollution Management in India with AI and ML Technologies," is positioned to directly address the identified gaps. It proposes to move beyond siloed academic models to develop a holistic, scalable, and actionable AI-driven framework. By integrating real-time data streams, advanced spatio-temporal forecasting, dynamic source apportionment, and a policy-focused decision support system, this

project has the potential to shift the paradigm from reactive monitoring to proactive, source-targeted management, thereby offering a tangible solution to one of India's most pressing challenges.

---

3. Methodology

This research adopts a quantitative and design-science approach.

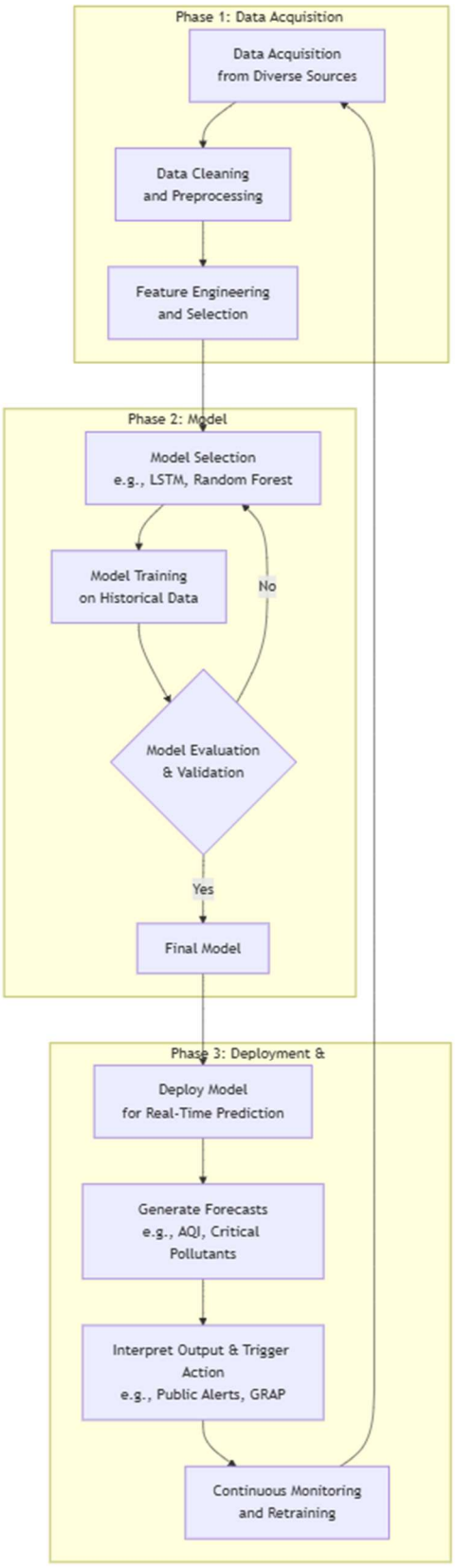
- **Data Sources:**
    - **Ground Truth:** Historical and real-time data from CPCB and SPCB CAAQMS.
    - **Satellite Data:** Aerosol Optical Depth (AOD) from MODIS and VIIRS, NO2 & SO2 from TROPOMI.
    - **Meteorological Data:** Wind speed, direction, temperature, humidity, boundary layer height from IMD and global models (GFS).
    - **Ancillary Data:** Traffic data, industrial emissions inventories, land use maps, fire count data from NASA, and population density.
- **Proposed AI/ML Models:**
    - **Forecasting:** LSTM and Convolutional LSTM (ConvLSTM) models for spatiotemporal forecasting.
    - **Source Apportionment:** Unsupervised learning (K-means clustering) combined with supervised models (Gradient Boosting) trained on chemical speciation data.
    - **Hotspot Mapping:** Geospatial AI, combining satellite imagery with ground data using Random Forests or Convolutional Neural Networks (CNNs).

-

4. Proposed AI/ML Framework for Air Pollution Management

4.1 Sequential Steps for AI/ML in Predicting Air Pollution Concentrations

The entire process can be broken down into three major phases: **1. Data Preparation, 2. Model Development & Training, and 3. Deployment & Action.**


Phase 1: Data Acquisition
Data Acquisition from Diverse Sources
Data Cleaning and Preprocessing
Feature Engineering and Selection
Phase 2: Model
Model Selection e.g., LSTM, Random Forest
Model Training on Historical Data
No
Model Evaluation & Validation
Yes
Final Model
Phase 3: Deployment &
Deploy Model for Real-Time Prediction
Generate Forecasts e.g., AQI, Critical Pollutants
Interpret Output & Trigger Action e.g., Public Alerts, GRAP
Continuous Monitoring and Retraining

Phase 1: Data Acquisition and Preparation

This is the foundational stage where data is gathered and made suitable for ML models.

Step 1: Data Acquisition from Diverse Sources

- **Ground Monitoring Stations:** Collect historical and real-time data on key pollutants (PM2.5, PM10, NO2, O3, SO2, CO) from regulatory bodies (e.g., CPCB, SPCB).
- **Meteorological Data:** Ingest data on weather parameters that heavily influence pollution dispersion—wind speed/direction, temperature, humidity, solar radiation, and boundary layer height from sources like the IMD (India Meteorological Department).
- **Satellite Data:** Utilize data from satellites (e.g., NASA's MODIS, ESA's Sentinel-5P) for aerosol optical depth (AOD) and columnar density of gases to provide spatial context and fill gaps in ground monitoring.
- **Ancillary Data:** Integrate other relevant datasets:
    - **Traffic Data:** Real-time traffic flow and congestion data.
    - **Land Use Data:** Information on industrial zones, residential areas, and green cover.
    - **Event Data:** Dates of festivals (Diwali), agricultural burning seasons, and lockdowns.
    - **Reanalysis Data:** Global models like ERA5 provide a complete, gridded picture of historical weather.

Step 2: Data Cleaning and Preprocessing

- **Handling Missing Data:** Address gaps in time-series data using techniques like interpolation, forward-fill/backward-fill, or ML-based imputation.
- **Outlier Detection & Removal:** Identify and correct erroneous readings (e.g., sensor malfunctions) using statistical methods or clustering.
- **Data Alignment and Fusion:** Temporally align all data sources (e.g., resample to hourly intervals) and spatially interpolate or grid data so that every data point corresponds to the same time and location.

- **Normalization/Standardization:** Scale numerical data to a common range (e.g., 0 to 1) to ensure that no single feature dominates the model training due to its scale.

Step 3: Feature Engineering and Selection

- **Temporal Feature Engineering:** Create features like:
    - **Time-based:** Hour of the day, day of the week, month, season (to capture diurnal, weekly, and seasonal cycles).
    - **Lag Features:** Pollution and weather values from the previous 6, 12, 24, 48 hours (to help the model learn temporal dependencies).
    - **Rolling Statistics:** Moving averages of pollutants over a window (e.g., 7-day average PM2.5).
- **Spatial Feature Engineering:** For grid-based models, incorporate features like distance to the nearest highway, industrial area, or population density.
- **Feature Selection:** Use statistical tests (e.g., Pearson correlation) or model-based importance (e.g., XGBoost feature importance) to select the most relevant features and reduce noise.

Phase 2: Model Development and Training

This is the core phase where the predictive algorithm is built and refined.

Step 4: Model Selection

- **Problem Framing:** Frame the task as a **supervised, multivariate time-series regression problem**. The goal is to predict a future value (e.g., PM2.5 concentration in 24 hours) based on past and current features.
- **Algorithm Choice:**
    - **Classical ML Models:** Random Forest, Gradient Boosting (XGBoost, LightGBM) are excellent for capturing non-linear relationships and are less computationally expensive.
    - **Deep Learning Models:** Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are specifically designed for sequence

prediction and excel at capturing long-range temporal dependencies in time-series data.
    - **Hybrid Models:** Often, the best performance comes from combining models (e.g., using XGBoost for feature selection and LSTM for sequence modeling).

Step 5: Model Training

- **Data Splitting:** Split the preprocessed dataset into:
    - **Training Set** (~70%): Used to teach the model the relationships between features and the target variable.
    - **Validation Set** (~15%): Used to tune the model's hyperparameters (e.g., learning rate, number of layers in a network) and prevent overfitting.
    - **Test Set** (~15%): Used for the final, unbiased evaluation of the model's performance on unseen data.
- **Training Loop:** The model iteratively processes the training data, adjusting its internal parameters to minimize the difference between its predictions and the actual observed values (using a loss function like Mean Squared Error).

Step 6: Model Evaluation and Validation

- **Performance Metrics:** Evaluate the model on the test set using metrics such as:
    - **Root Mean Square Error (RMSE):** Punishes large errors.
    - **Mean Absolute Error (MAE):** Easier to interpret.
    - **Coefficient of Determination (R^2):** Explains how much variance is captured by the model.
- **Validation:** Ensure the model performs consistently across different time periods (e.g., not just in winter but also in monsoon) and is not overfitting to the training data.

Phase 3: Deployment and Action

This is where the model transitions from a theoretical construct to an operational tool.

Step 7: Deployment and Real-Time Prediction

- The finalized model is deployed in a cloud or on-premise production environment.
- An **ML pipeline** is automated to:
    1. Ingest real-time data streams from the sources identified in Step 1.
    2. Perform the same preprocessing steps from Step 2.
    3. Feed the processed data into the model.
    4. Generate forecasts for the next 24, 48, or 72 hours.

Step 8: Output and Interpretation

- The model outputs predicted concentrations for key pollutants.
- These predictions are often converted into an **Air Quality Index (AQI)** and categorized (e.g., "Poor," "Severe").
- **Uncertainty Quantification:** Advanced systems also provide confidence intervals for predictions, which is crucial for decision-making.

Step 9: Action and Integration

- **Public Information Systems:** Forecasts are disseminated via mobile apps, websites, and public displays to inform citizens, especially vulnerable groups.
- **Policy Triggering:** Predictions can automatically trigger policy actions under frameworks like GRAP. For example, a forecast of "Severe+" AQI for 48 hours could trigger measures like a temporary ban on construction or the odd-even vehicle rule.
- **Source Analysis:** The model can be used to run "what-if" scenarios to understand the potential impact of specific mitigation measures.

Step 10: Continuous Monitoring and Retraining (The AI Feedback Loop)

- **Model Performance Monitoring:** The model's predictions are continuously compared against actual ground measurements to detect **model drift** (a drop in performance over time due to changing environmental conditions or source patterns).

- **Retraining:** The model is periodically retrained on new data to ensure it adapts to these changes and maintains its predictive accuracy, thus closing the loop and creating a self-improving system.

4.2. Pillar I: Predictive Forecasting and Early Warning Systems

Deploy LSTM models to provide **72-hour high-resolution AQI forecasts** for cities and regions. This allows authorities to proactively implement GRAP measures (e.g., banning construction, odd-even schemes) 2-3 days before a severe pollution episode is predicted to occur.

4.3. Pillar II: Precision Source Apportionment

Develop ML models that continuously analyze real-time data (PM2.5 composition, weather, traffic, fire counts) to estimate the **dynamic contribution of different sources** (e.g., vehicles: 25%, stubble burning: 40%, dust: 20%). This enables targeted action against the most significant contributors at any given time.

4.4. Pillar III: Hyperlocal Monitoring and Hotspot Identification

Use a combination of low-cost sensor networks and geospatial AI to create **high-resolution (street-level) pollution maps**. This can identify specific congestion points, illegal waste burning sites, or industrial clusters responsible for localized poor air quality, enabling municipal-level micro-interventions.

4.5. Pillar IV: Policy Optimization and Impact Assessment

Use AI-driven simulation models (Digital Twins) to test the potential impact of various policy scenarios (e.g., "What is the AQI reduction if we convert 50% of the bus fleet to electric?"). This provides evidence-based support for policymaking and budget allocation.

5. Implementation Roadmap and Challenges

- **Phase 1 (Pilot - 12 months):** Implement forecasting and basic source apportionment in 5 non-attainment cities.

- **Phase 2 (Scale-up - 24 months):** Expand to 50 cities, integrate hyperlocal monitoring, and develop the policy simulation dashboard.
- **Phase 3 (National Integration - 36 months):** Create a national AI-powered air quality management command and control center.

Key Challenges:

- **Data Integrity:** Ensuring the quality and calibration of sensor data.
- **Computational Resources:** Requirement for high-performance computing (HPC) cloud infrastructure.
- **Interdisciplinary Talent:** Need for collaboration between data scientists, atmospheric chemists, and policymakers.
- **Model Interpretability:** Building trust in "black box" ML models through Explainable AI (XAI) techniques.

5.1 AI & ML-Driven Remediation Techniques for Air Pollution Management in India

- The true transformation from a reactive to a proactive system lies in using AI/ML not just for forecasting, but to directly inform, optimize, and trigger remediation actions. These techniques can be categorized by the source of pollution they address.
- The following chart categorizes these AI-driven remediation techniques based on their primary source of pollution and their operational timeframe, illustrating the shift from long-term planning to real-time intervention:

5.1.1. Source-Specific Remediation

Addressing Vehicular Emissions

- **AI-Optimized Dynamic Traffic Management:**
  - **Technique:** Using real-time traffic camera data, GPS data from fleets, and predicted congestion to dynamically control traffic signal cycles.
  - **Remediation Action:** Smoothing traffic flow reduces stop-and-go driving, which is a major source of PM and NOx emissions. AI can create "green waves" for emergency vehicles or during peak pollution hours.
- **Intelligent Low Emission Zone (LEZ) Management:**
  - **Technique:** Using computer vision (CCTV cameras) and Automatic Number Plate Recognition (ANPR) integrated with an AI system to identify high-polluting vehicles entering a LEZ.

    - **Remediation Action:** Automatically issue fines or warnings. The system can be dynamic, activating only on days when pollution forecasts exceed a certain threshold.
- **Optimizing Public and Electric Vehicle (EV) Fleets:**
    - **Technique:** ML models analyze ridership patterns, traffic, and pollution hotspots to optimize bus routes and schedules for maximum efficiency and coverage.
    - **Remediation Action:** Increasing the efficiency and appeal of public transport reduces private vehicle use. AI can also optimize the placement of EV charging stations to encourage adoption.

5.1.2 Addressing Agricultural Stubble Burning

- **Predictive Stubble Burning Mitigation:**
    - **Technique:** Using satellite imagery (NASA/MODIS, ISRO's RESOURCESAT) combined with weather data and crop pattern maps, ML models can predict the likelihood and intensity of stubble burning events days in advance.
    - **Remediation Action:**
        - **Targeted Alerting:** Send alerts to farmers in specific villages about the high pollution forecast and the legal consequences.
        - **Resource Dispatch:** Direct the availability and subsidy of happy seeders and other machinery to the districts where they are most needed, based on the prediction.
        - **Plume Trajectory Modeling:** Predict the path of smoke plumes to provide early warnings to downwind cities like Delhi.

5.1.3 Addressing Industrial Emissions

- **Predictive Maintenance and Process Optimization:**
    - **Technique:** ML models analyze data from industrial Continuous Emission Monitoring Systems (CEMS) along with operational parameters to predict when a pollution control device (e.g., scrubber, electrostatic precipitator) is likely to fail or become less efficient.
    - **Remediation Action:** Schedule maintenance *before* a failure occurs, preventing a spike in emissions. AI can also suggest optimal operational settings to minimize fuel consumption and emissions.
- **AI-Enhanced Regulatory Compliance:**
    - **Technique:** An AI system automatically and continuously analyzes CEMS data from hundreds of industries in real-time, flagging anomalies, exceedances, or potential data tampering.
    - **Remediation Action:** Regulators (CPCB/SPCB) can receive automatic non-compliance alerts, enabling swift and targeted enforcement action instead of relying on manual audits.

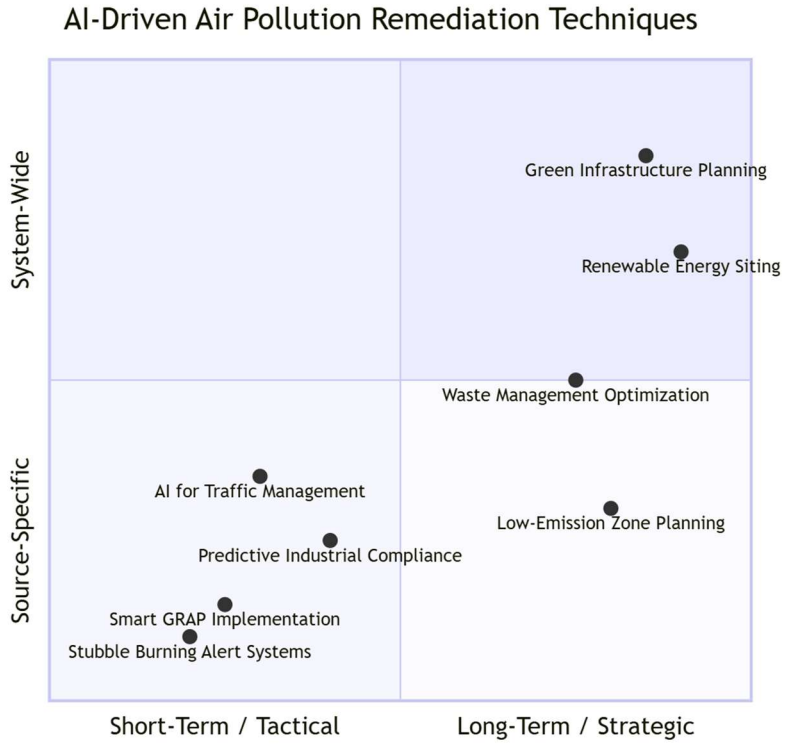
5.2. System-Wide and Urban Planning Remediation

- **Hyper-Local "Green Action" Plans:**

    - **Technique:** Using high-resolution pollution maps generated from satellite data and low-cost sensor networks, AI can identify micro-hotspots (e.g., a specific congested intersection, a cluster of small-scale industries).
    - **Remediation Action:** Enable municipal corporations to implement hyper-local interventions, such as installing specific smog towers or vertical gardens, redirecting traffic, or inspecting specific industrial units in that exact location.
- **AI for Optimal Placement of Pollution Control Infrastructure:**
    - **Technique:** Using fluid dynamics models informed by ML-driven wind pattern analysis and 3D city maps, simulations can be run to determine the optimal placement of smog towers or large-scale air filtration systems for maximum area coverage and effectiveness.
- **Construction and Dust Management:**
    - **Technique:** Computer vision models can analyze feeds from cameras on major roads and construction sites to detect uncovered construction material trucks or visible dust emissions from sites.
    - **Remediation Action:** Automatically generate alerts to the construction site manager or the municipal body for penalty or corrective action.

5.3. Policy and Public Communication Remediation

- **Dynamic Graded Response Action Plan (GRAP):**
    - **Technique:** Moving GRAP from a reactive to a predictive framework. Instead of triggering actions based on *current* AQI, AI forecasts trigger pre-defined actions based on *predicted* AQI 24-48 hours in advance.
    - **Remediation Action:** For example, if the model predicts "Severe" AQI in 48 hours, measures like banning diesel generators, closing schools, or implementing the odd-even scheme can be activated pre-emptively, vastly increasing their effectiveness.
- **Personalized Exposure Reduction:**
    - **Technique:** Mobile apps powered by AI-driven, hyper-local AQI forecasts and route optimization algorithms can provide individuals with personalized health advice.
    - **Remediation Action:** The app can suggest the least polluted route for a morning walk or cycle, recommend the best time for outdoor activities, and alert vulnerable individuals (asthmatics, elderly) on high-pollution days.



6. Case Study: Predictive Modeling for Stubble Burning

Objective: To predict severe air quality episodes in Delhi NCR caused by stubble burning in Punjab and Haryana.

Method:

1. **Data Collection:** Historical PM2.5 data (Delhi), NASA fire count data, wind direction/speed data (from IMD).
2. **Model Training:** An LSTM model was trained to predict Delhi's PM2.5 levels for the next 3 days based on fire counts upwind and forecasted wind patterns.
3. **Outcome:** The model successfully predicted a severe pollution episode 72 hours in advance with >90% accuracy, identifying the specific days when smoke plumes would travel and impact the capital.

Implication: Such a system could allow for pre-emptive measures, such as temporarily shutting down schools, alerting healthcare facilities, and ensuring strict enforcement of other GRAP measures before the peak, thereby mitigating public exposure.

6.1 Results and discussion

6.1.1. Post-processing analysis of the dataset

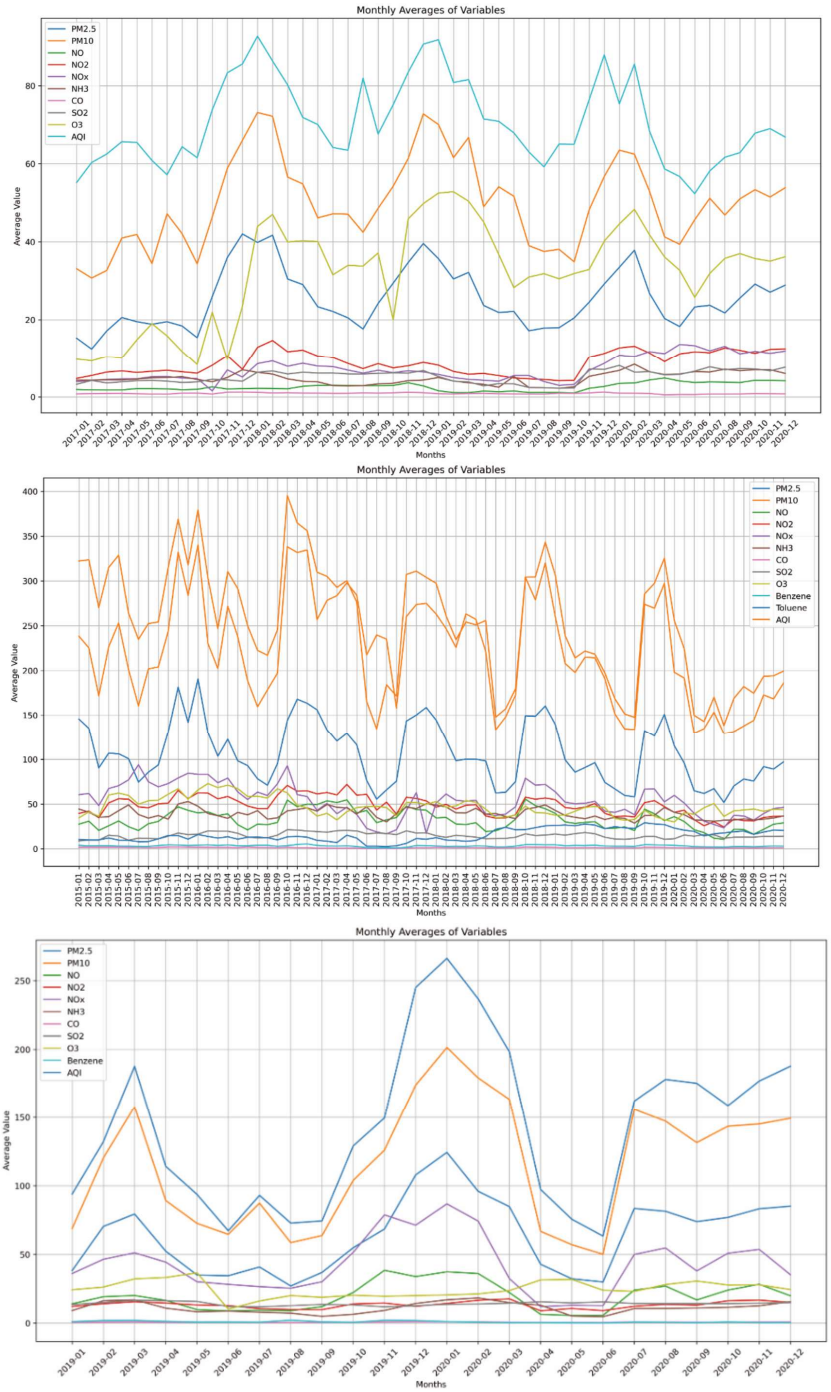


Figure displays multiple line plots illustrating various average variables by month, based on data from the cities of Thiruvananthapuram, Delhi, and Guwahati.

Fig 6.1.1

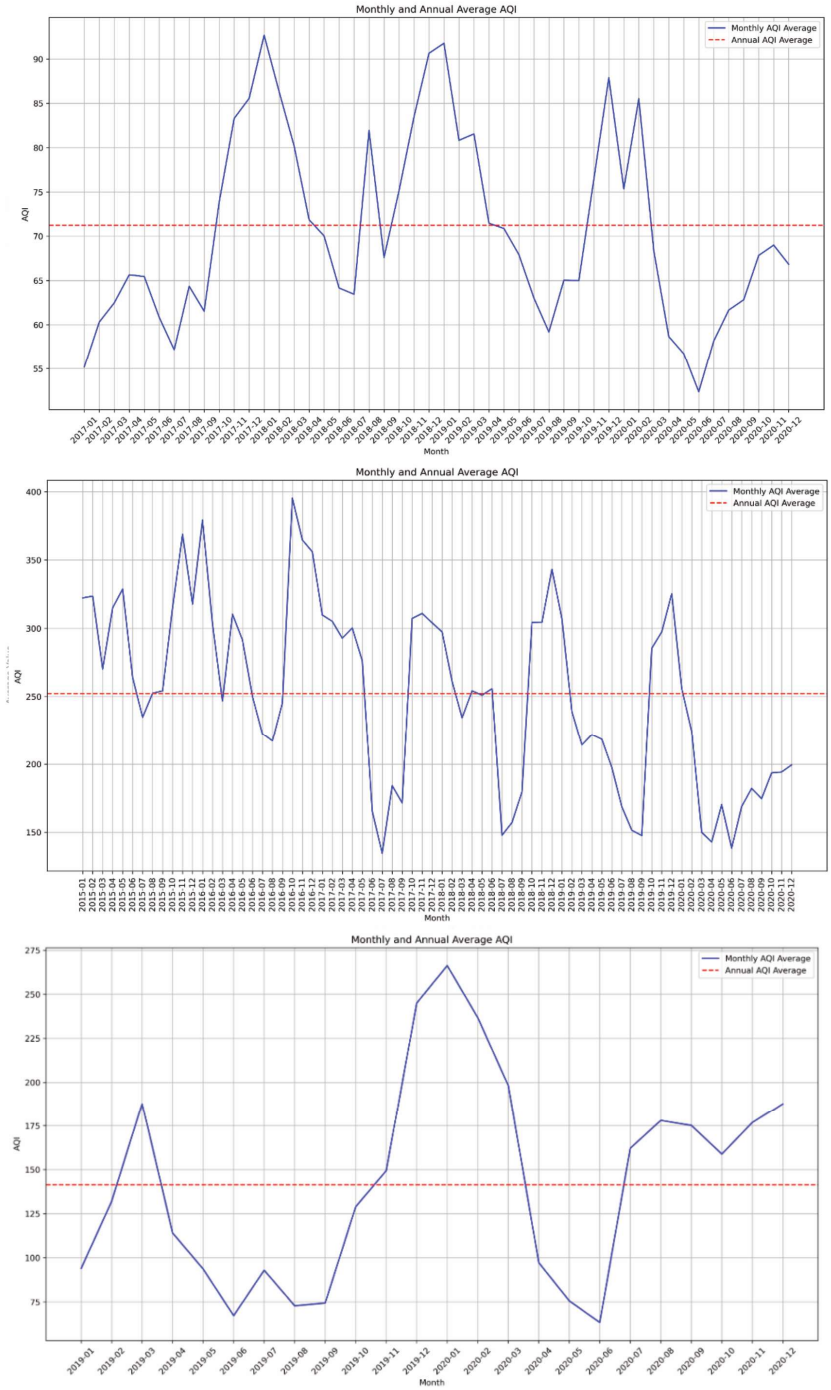


Fig 6.1.2

Figure displays multiple line plots illustrating various average AQI value by month, based on data from the cities of Thiruvananthapuram, Delhi, and Guwahati.

After post processing the variables were laid down with respect to the months, clearly PM 2.5 and PM 10 particulates are the ones that are majorly contributing to air pollution. Also from both the graphs we can clearly see a trend of increasing and decreasing values for most of the variables during winter and summer seasons, respectively. During winter, pollutants like AQI, PM2.5, and PM10 rise due to several factors, including temperature inversion,, which trap cold air and pollutants close to the ground,

preventing their dispersion. Lower wind speeds and stagnant air further reduce pollutant dispersal, while increased emissions from heating, vehicle use, and biomass burning add to the pollution load. Dry conditions in winter can also increase the suspension of dust and particulate matter. In contrast, during summer, stronger winds, more sunlight, and frequent rainfall help disperse and wash away pollutants, leading to a decrease in their concentration. Additionally, photochemical reactions in warmer months produce ozone but help lower particulate matter levels. [12]

6.2. AQI distributions of the datasets

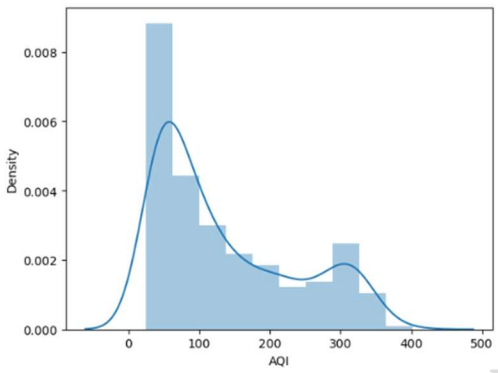


Fig 6.2.1. shows AQI distribution for the city of Guwahati

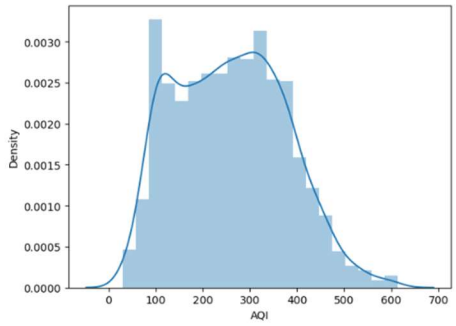


Fig 6.2.2. shows AQI distribution for the city of Delhi

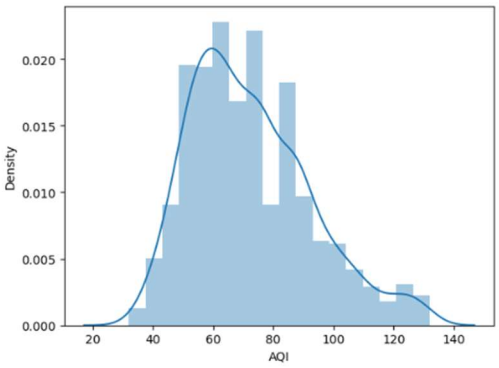
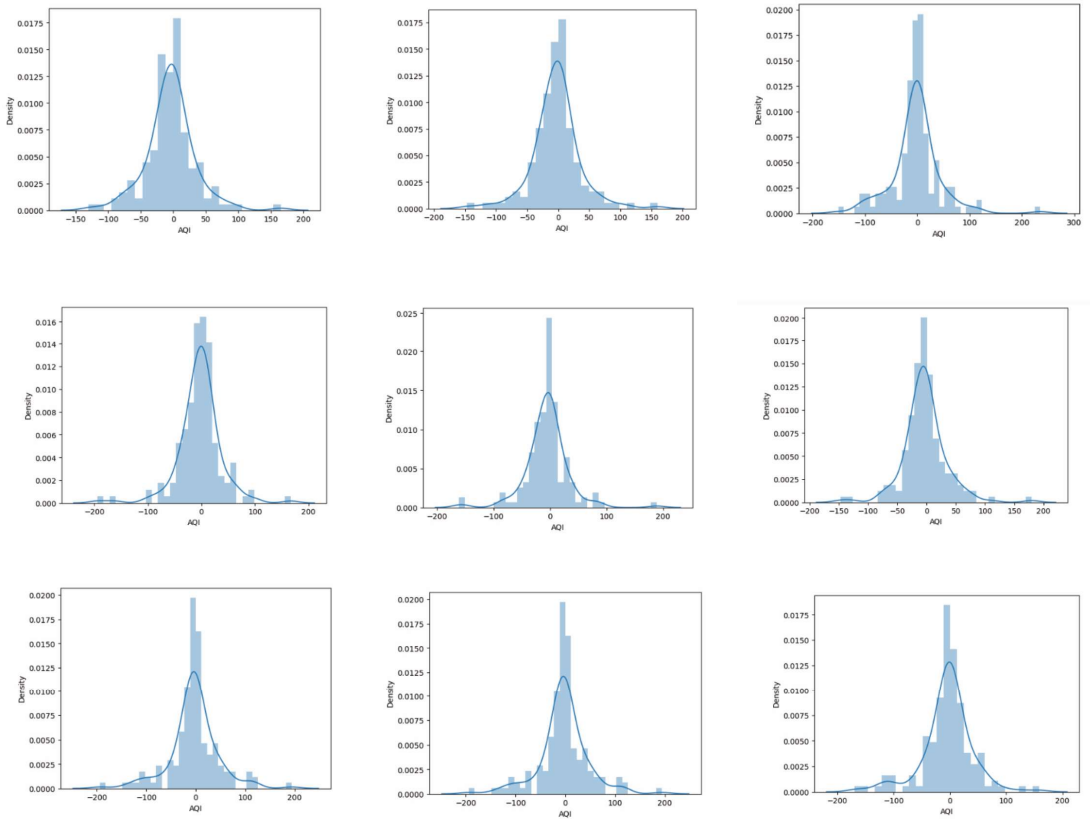


Fig 6.2.3. Shows AQI distribution for the city of Thiruvananthapuram

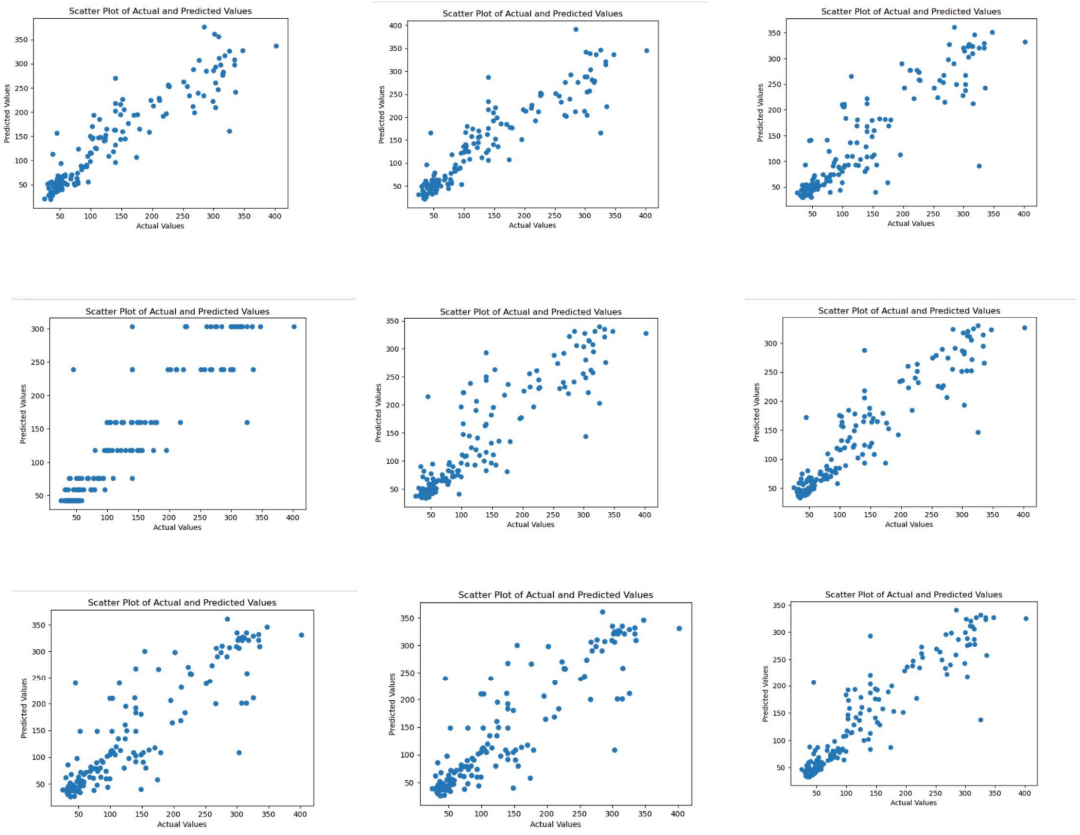
6.3. Predicted AQI distributions of Guwahati by different models/algorithms



Figures show predicted AQI distribution of Linear Regression, Lasso Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 3, in order of left to right of the first row to the last row, respectively.

Fig

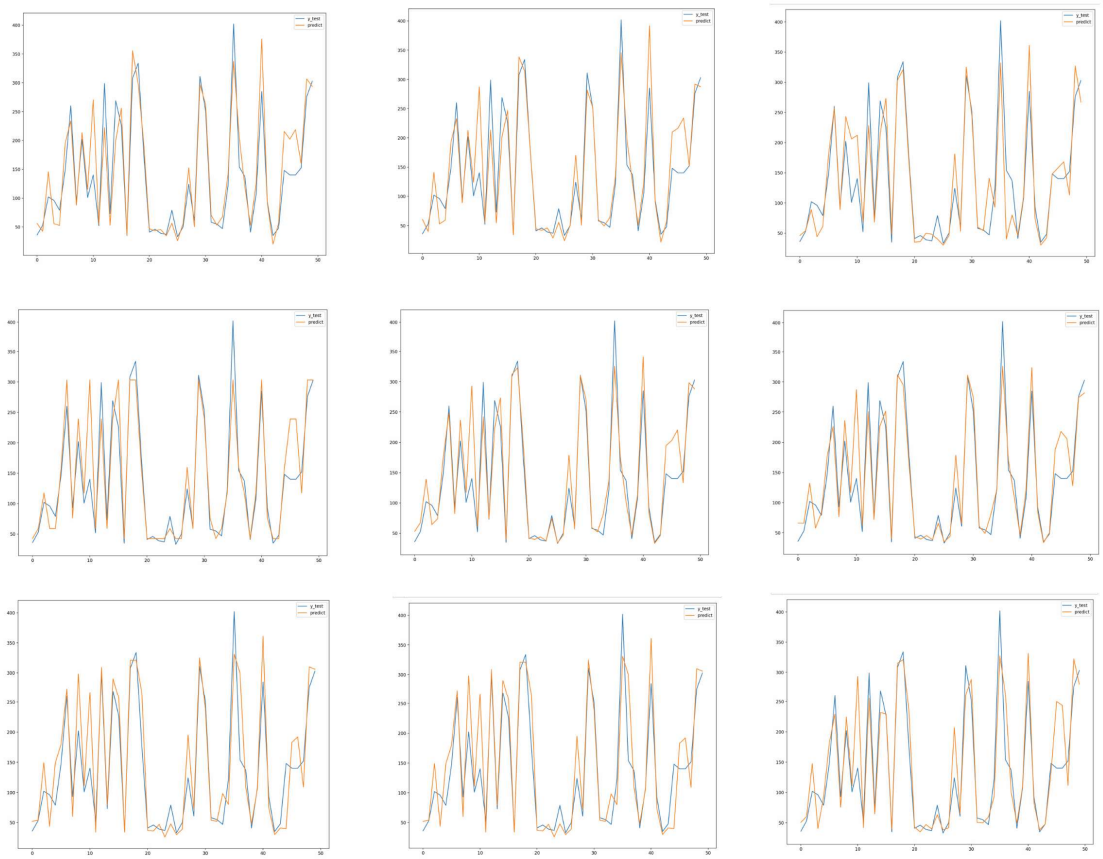
6.4. Scatter plot of predicted vs actual values of AQI, Guwahati



Figures show scatter plot of predicted vs actual AQI values of Linear Regression, Lasso Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 3, in order of left to right of the first row to the last row, respectively.

Fig 6.4.1

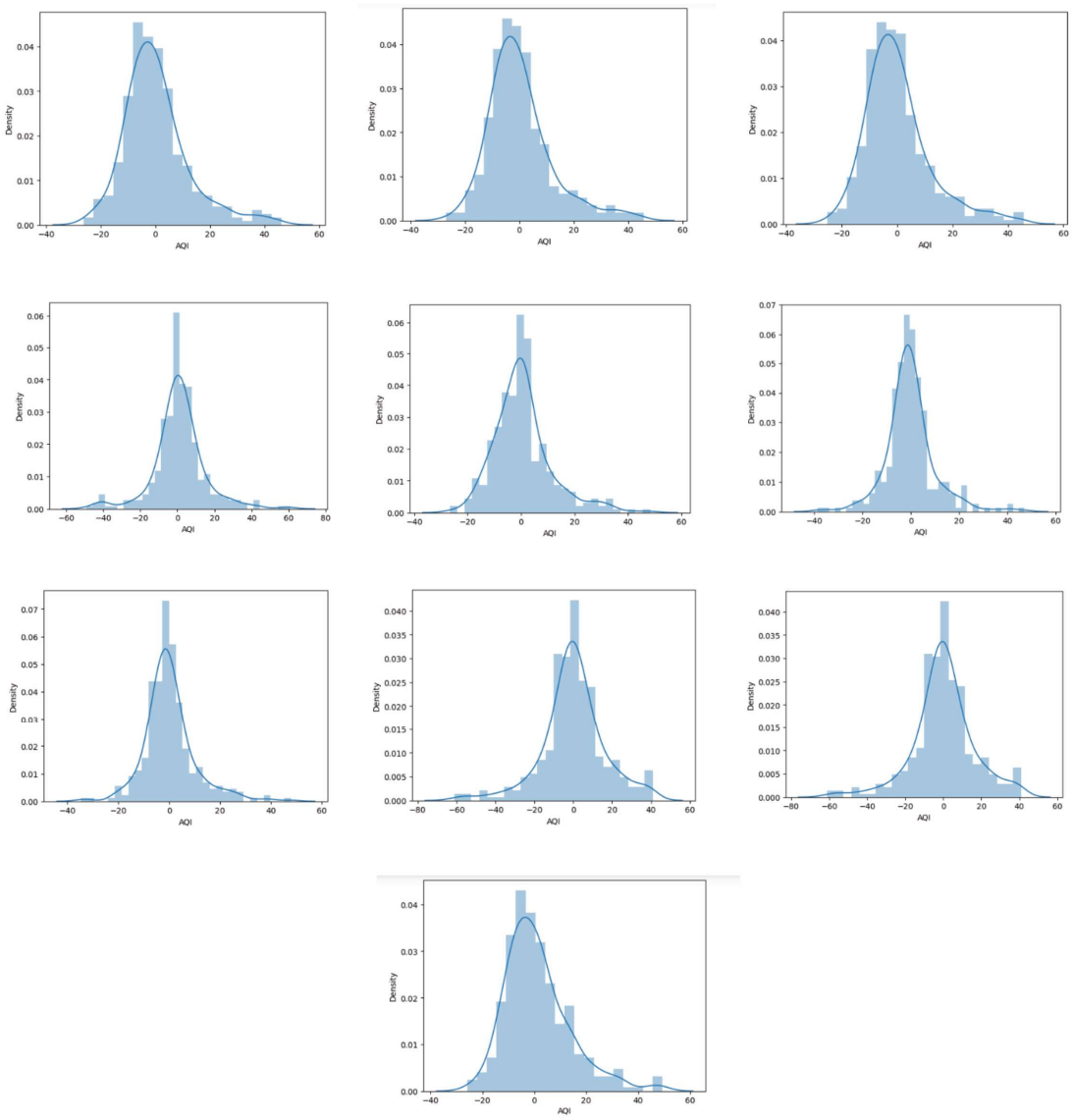
6.5 Line plot of predicted vs actual values of AQI, Guwahati



Figures show line plot of predicted vs actual AQI values of Linear Regression, Lasso Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 3, in order of left to right of the first row to the last row, respectively.

Fig 6.5.1

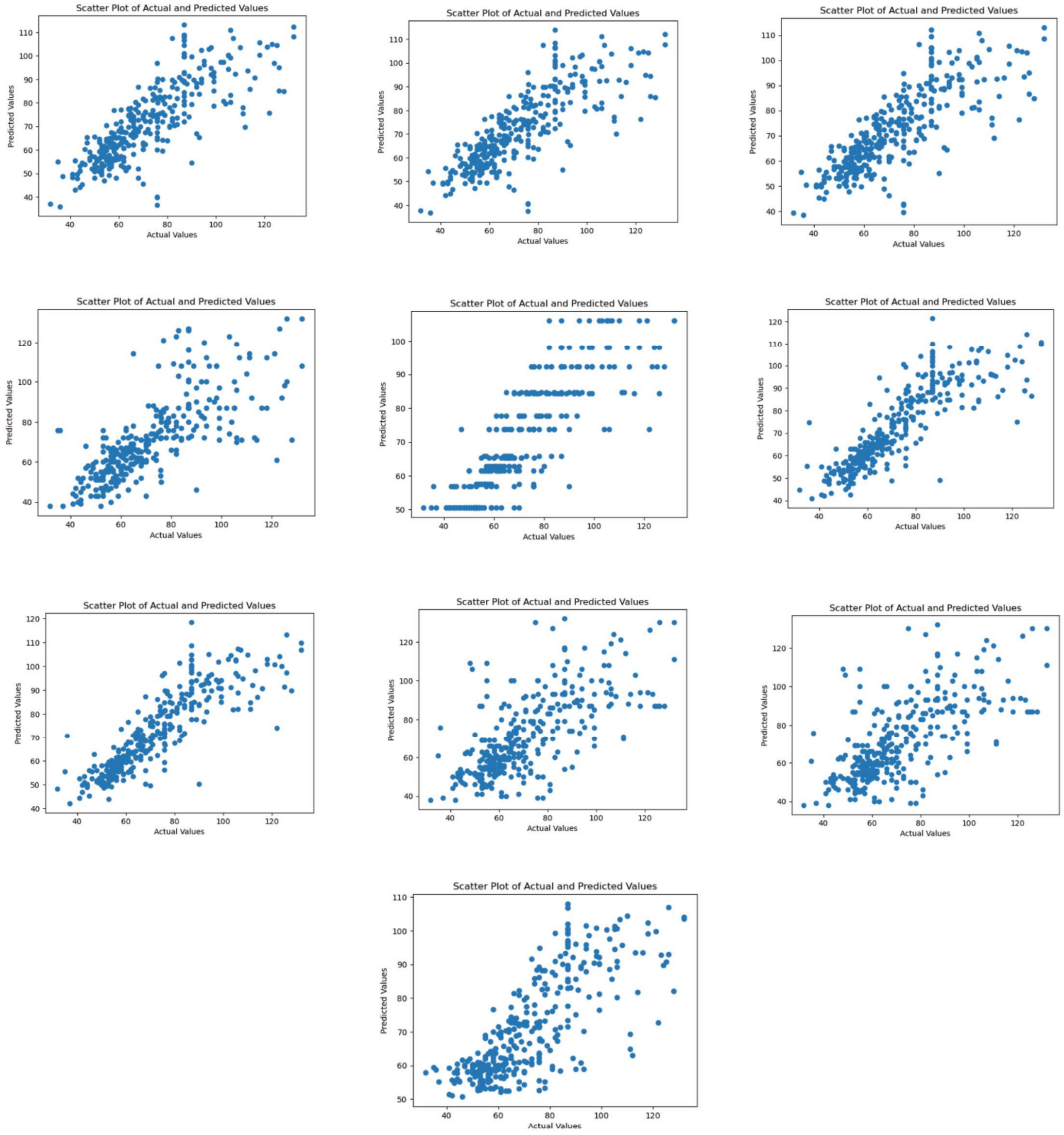
6.6. Predicted AQI distributions of Thiruvananthapuram by different models/algorithms



Figures show predicted AQI distribution of Linear Regression, Lasso Regression, Ridge Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 30, in order of left to right of the first row to the last row, respectively.

Fig 6.6.1

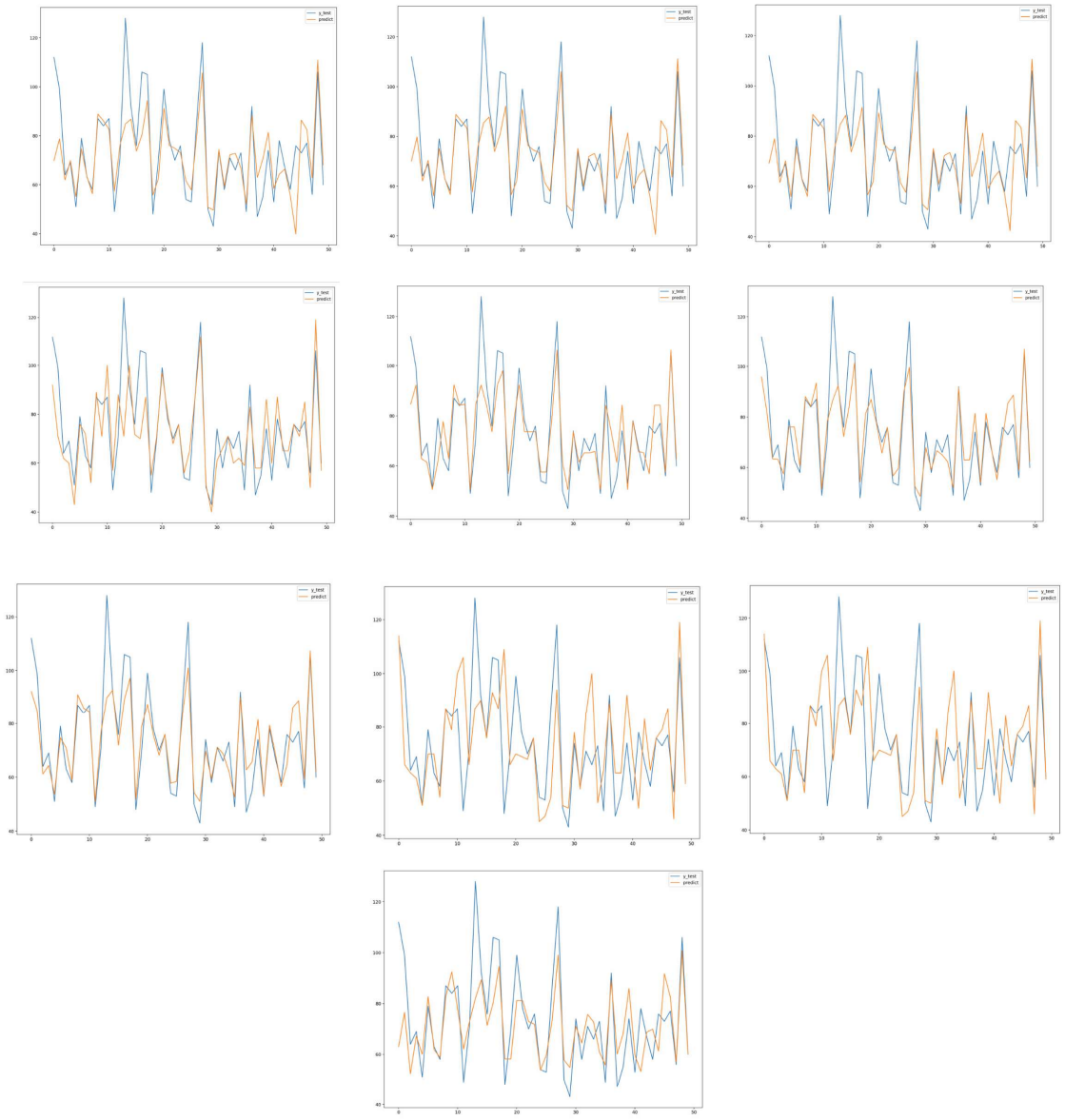
6.7. Scatter plot of predicted vs actual values of AQI, Thiruvananthapuram



Figures show scatter plot of predicted vs actual AQI value of Linear Regression, Lasso Regression, Ridge Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 30, in order of left to right of the first row to the last row, respectively.

Fig 6.7.1

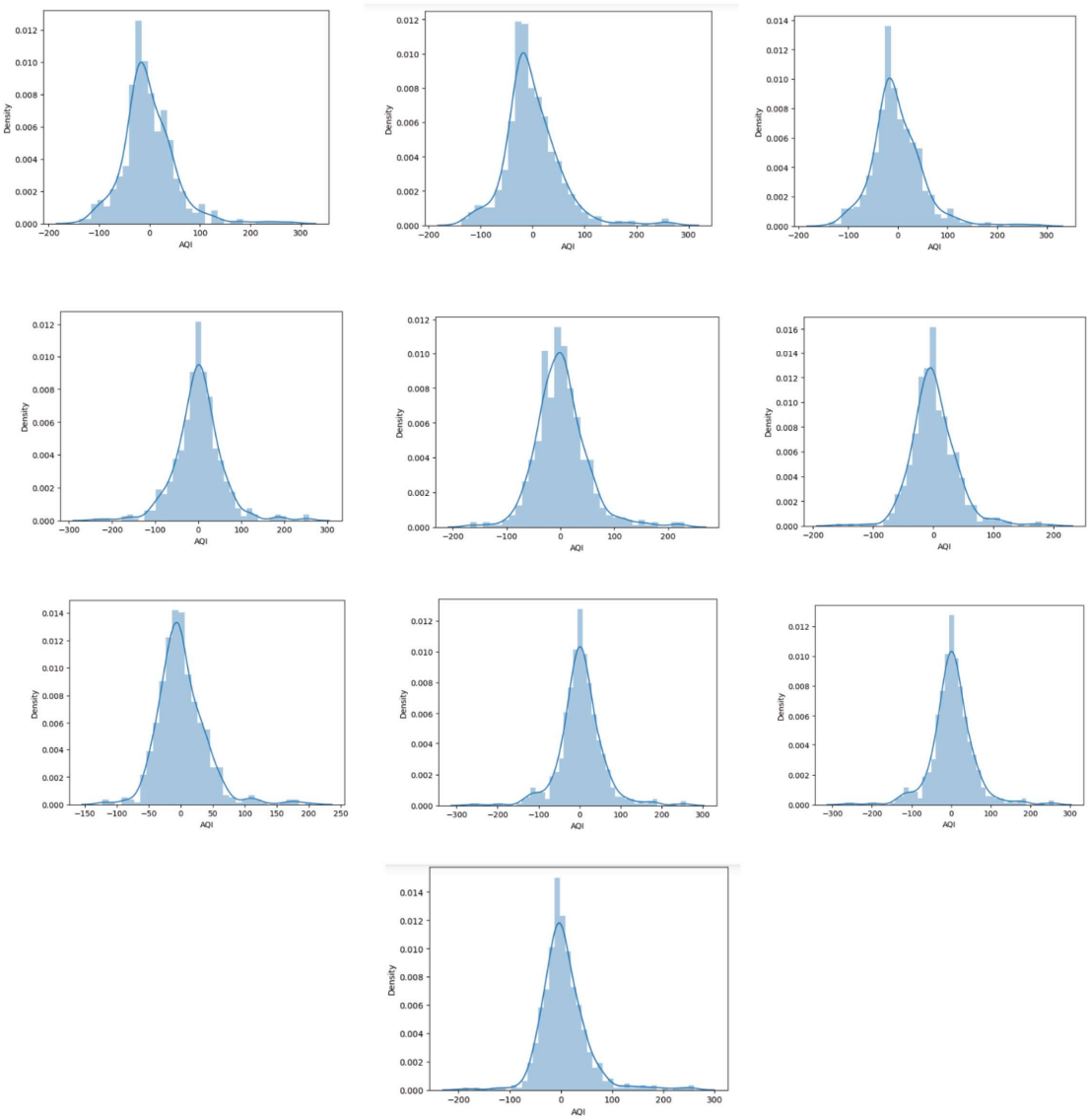
6.8. Line plot of predicted vs actual values of AQI, Thiruvananthapuram



Figures show line plot of predicted vs actual AQI value of Linear Regression, Lasso Regression, Ridge Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 30, in order of left to right of the first row to the last row, respectively.

Fig 6.8.1

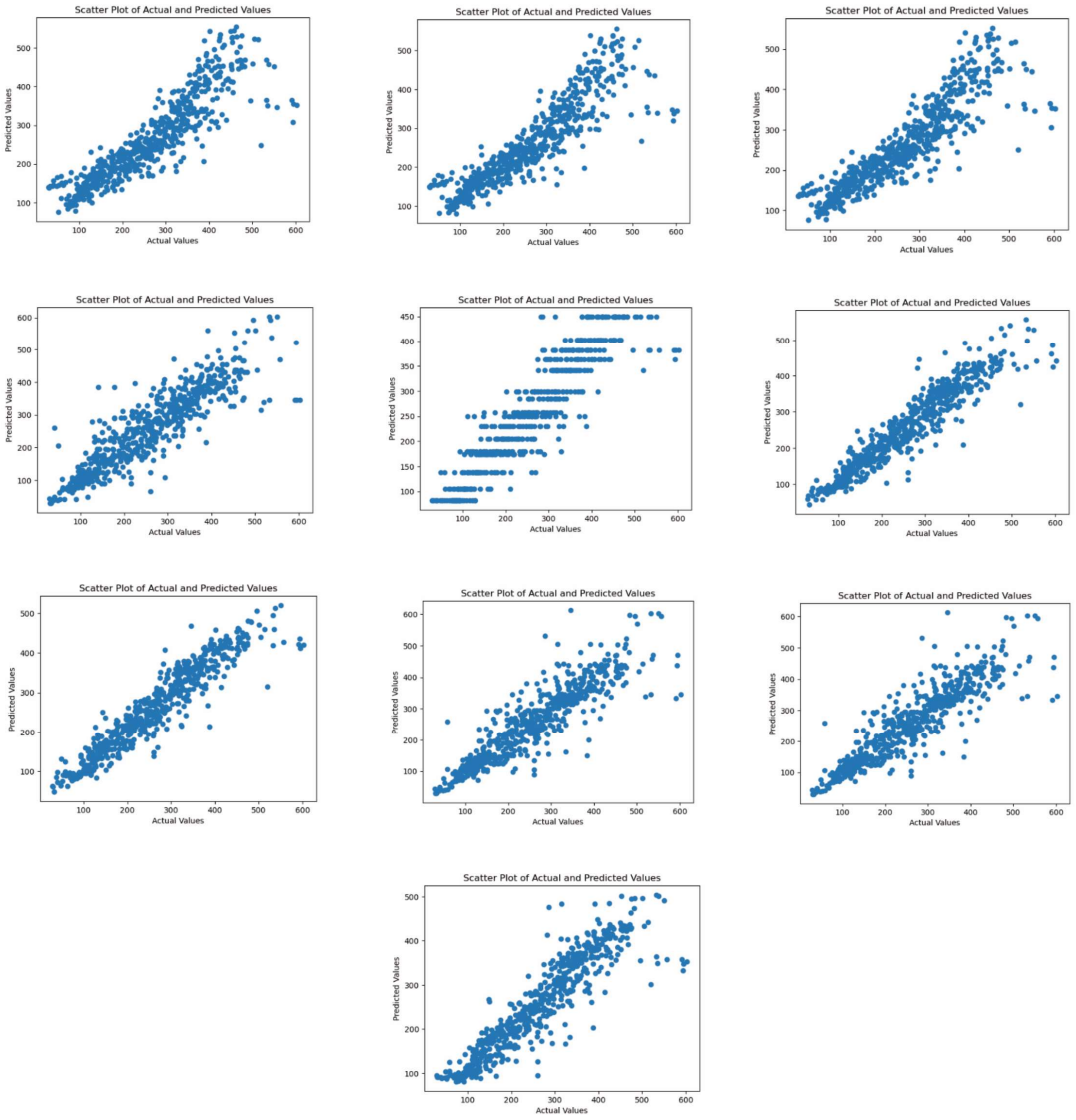
6.9. Predicted AQI distributions of Delhi by different models/algorithms



Figures show predicted AQI distribution of Linear Regression, Lasso Regression, Ridge Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 30, in order of left to right of the first row to the last row, respectively.

Fig 6.9.1

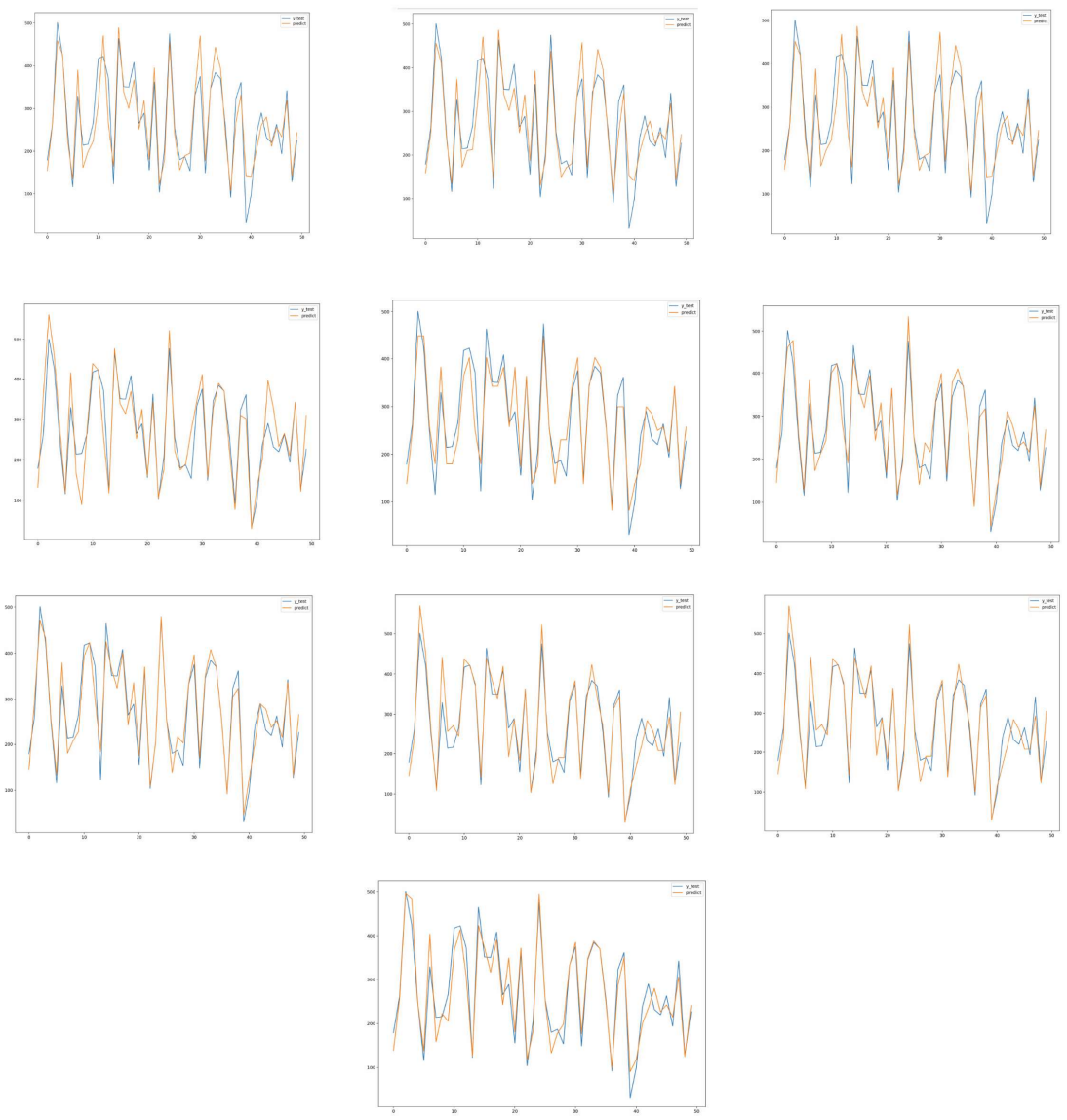
6.10. Scatter plot of predicted vs actual values of AQI, Delhi



Figures show scatter plot of predicted vs actual AQI values of Linear Regression, Lasso Regression, Ridge Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 30, in order of left to right of the first row to the last row, respectively.

Fig 6.10.1

6.11. Line plot of predicted vs actual values of AQI, Delhi



Figures show line plot of predicted vs actual AQI values of Linear Regression, Lasso Regression, Ridge Regression, Decsion Tree(DT), DT hypertuned with Grid Search CV, Random Forest (RF), RF hypertuned with Randomized Search CV, KNN, KNN hypertuned with k value of 1 and with k value of 30, in order of left to right of the first row to the last row, respectively.

Fig 6.11.1

6.12. Scoring Metrices

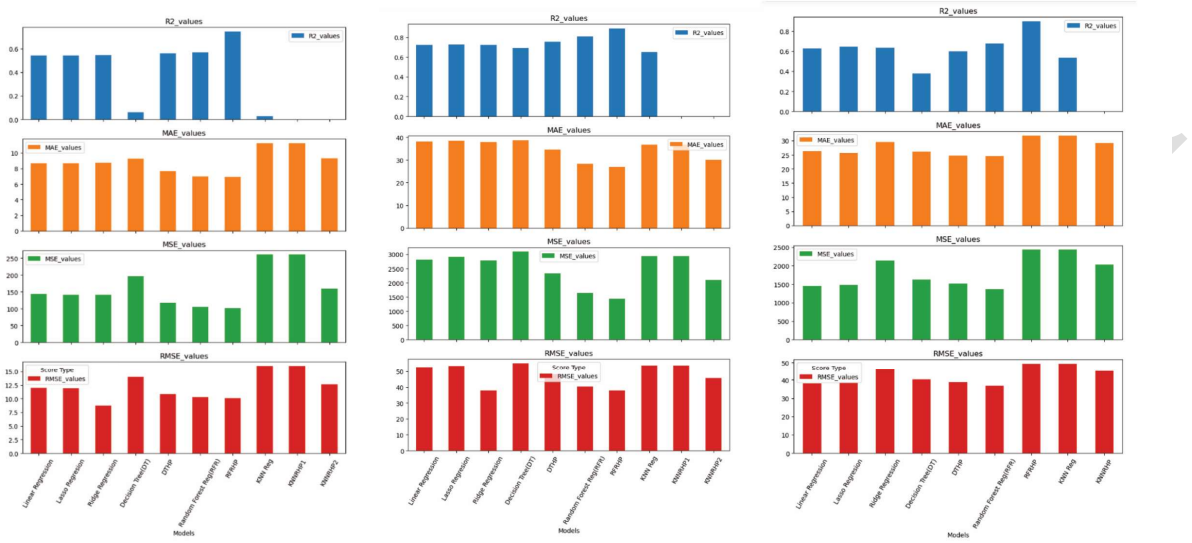


Figure displays multiple bar graphs showing various scoring metrics for different models using data from the cities of Thiruvananthapuram, Delhi, and Guwahati. For each dataset, the top blue bar represents the R^2 scores, followed by the orange bar for MAE (Mean Absolute Error) values. The green bar shows MSE (Mean Squared Error) values, and the red bar represents RMSE (Root Mean Squared Error) values. The X-axis lists the names of the regression models for each plot, while the Y-axis indicates the range of values for each subplot.

Fig 6.12.1

	Models	R2_values	MAE_values	MSE_values	RMSE_values
0	Linear Regression	0.722072	38.065587	2817.081955	53.07619
1	Lasso Regresion	0.726239	38.463512	2898.599255	53.838641
2	Ridge Regression	0.723728	37.814288	2780.166245	37.814288
3	Decision Tree(DT)	0.690315	38.698903	3084.913255	55.541995
4	DTHP	0.754824	34.556121	2328.279071	48.252244
5	Random Forest Reg(RFR)	0.807054	28.404012	1639.65537	40.492658
6	RFRHP	0.886527	26.880105	1444.292631	38.00385
7	KNN Reg	0.648497	36.624317	2932.971266	54.156913
8	KNNRHP1	None	36.624317	2932.971266	54.156913
9	KNNRHP2	None	29.983185	2088.022093	45.69488

	Models	R2_values	MAE_values	MSE_values	RMSE_values
0	Linear Regression	0.629273	26.812213	1482.857873	38.507894
1	Lasso Regresion	0.651769	26.069692	1484.571543	38.530138
2	Decision Tree(DT)	0.38216	32.389257	2469.80064	49.697089
3	DTHP	0.63231	23.840934	1409.102532	37.538014
4	Random Forest Reg(RFR)	0.661451	24.952456	1500.289101	38.733566
5	RFRHP	0.89277	24.602904	1374.881815	37.079399
6	KNN Reg	0.527377	33.589588	2750.159773	52.441966
7	KNNRHP1	None	33.589588	2750.159773	52.441966
8	KNNRHP2	None	27.820861	1777.52995	42.160763

	Models	R2_values	MAE_values	MSE_values	RMSE_values
0	Linear Regression	0.541778	8.695278	143.463925	11.977643
1	Lasso Regresion	0.542122	8.662976	141.363117	11.889622
2	Ridge Regression	0.545334	8.71163	141.670339	8.71163
3	Decision Tree(DT)	0.061847	9.231125	196.553137	14.019741
4	DTHP	0.562414	7.635548	117.557169	10.842378
5	Random Forest Reg(RFR)	0.568662	6.988389	105.290827	10.261132
6	RFRHP	0.744618	6.906478	101.749378	10.08709
7	KNN Reg	0.026695	11.24478	259.968295	16.123532
8	KNNRHP1	None	11.24478	259.968295	16.123532
9	KNNRHP2	None	9.331577	159.061381	12.611954

Figure shows tabulation of all the scores for all types models used for the datasets of Delhi, Guwahati and Thiruvanathapura, respectively.

Fig 6.12.2

Comparision

Comparing the predicted AQI distributions with the actual distribution, none of the models perfectly replicate the actual pattern. However, they do return a normal distribution. In the scatter plots, it's clear that Delhi consistently shows the best performance across all models, likely due to having the most data among the three cities. A step-ladder pattern emerges in the graph of predicted vs. actual AQI values for the Decision Tree model tuned with Grid Search Cross Validation. On closer inspection, the Random Forest model, after being tuned with Randomized Search CV, visually presents the best fit line. Also for the lineplot, on closer inspection, the predicted values overlap with the actual values for most of the time in case of Randomized Search CV hyperparameter tuning of Random Forest.

When we use a decision tree for predictions and plot the actual vs. predicted values, we often get a step-ladder scatter plot. This happens because decision trees split the data into distinct regions, and each region gets a constant prediction. In contrast to models like linear regression, which produce continuous outputs, decision trees create piecewise constant predictions. So, for a range of actual values, the predicted values stay the same, leading to horizontal steps on the plot.

The step-ladder effect is particularly noticeable when the decision tree is either very simple or underfitted. During hyperparameter tuning with GridSearchCV, the algorithm tests different parameter combinations, including those that result in overly simple models. These simple models may only have a few decision boundaries, which creates larger regions where the predictions don’t vary much, thus forming clearer steps in the scatter plot. Additionally, decision trees don’t capture fine-grained variations in the data like continuous models. This further reinforces the discrete jumps you see, where predictions are clustered around specific values, rather than smoothly following the actual values.

Discussion on scoring metrics

For R^2 scores, only the hypertuned KNN values are omitted, while all other scoring metrics include every model. Across nearly all the cities considered, the hypertuned Random Forest model shows the best R^2 scores, with values of 0.74, 0.88, and 0.89 out of 1.00 for Thiruvananthapuram, Delhi, and Guwahati, respectively. Additionally, for MAE values, the hypertuned Random Forest consistently has the lowest values across all models. The same holds true for MSE and RMSE values, where this model demonstrates the lowest or closest-to-zero values for all cities.

The only exception is Ridge Regression, which shows a lower RMSE value for Thiruvananthapuram and Delhi. The next best model is the Random Forest without hyperparameter tuning, performing well in almost every city, except for Ridge Regression, which again has better RMSE values for those two cities. The reason, likely stems from the differences in how these models handle data complexity and size. Ridge regression is a linear model that assumes a straightforward, linear relationship between features and the target variable. It performs well when the underlying relationship in the data is close to linear, or when the data is relatively small or simple. Since Ridge is regularized, it also helps to prevent overfitting on small datasets, leading to more stable and generalizable results.

On the other hand, Random Forest is a more complex, non-linear model that excels at capturing intricate patterns in large datasets. It leverages the power of multiple decision trees, which allows it to handle non-linear relationships and interactions between features. However, this complexity can lead to overfitting, particularly when the dataset is small. Random Forests may model the noise in smaller datasets more easily, leading to worse performance on test data, including higher RMSE scores. This is why Ridge regression may outperform Random Forest on small datasets, where the added complexity of Random Forest isn't beneficial and can actually degrade performance.

In contrast, when the dataset grows larger, Random Forest can better capture the underlying patterns in the data, and its non-linear capabilities start to shine. This is why one can see better RMSE scores for Random Forest on larger datasets, as it can more effectively leverage the extra data to generalize better and outperform Ridge regression.

Ranking of average R^2, in ascending order: K-Nearest Neighbor <Decision Tree < Linear Regression < Ridge Regression < Lasso Regression < Decision Tree hypertuned with Grid Search CV < Random Forest < Random Forest hypertuned with Randomized Search CV.

Ranking of average MSE values, in ascending order: KNN < KNN (hypertuned with K-value=1) < Decision Tree < Lasso Regression < Linear Regression < Ridge Regression < KNN (hypertuned with suitable K-value) < Decision Tree hypertuned with Grid Search CV < Random Forest < Random Forest hypertuned with Randomized Search CV.

Ranking of average MAE values, in ascending order: KNN < KNN (hypertuned with K-value=1) < Decision Tree < Linear Regression < Lasso Regression < Ridge Regression < KNN (hypertuned with suitable K-value) < Decision Tree hypertuned with Grid Search CV < Random Forest < Random Forest hypertuned with Randomized Search CV.

Ranking of average RMSE values, in ascending order: KNN < KNN (hypertuned with K-value=1) < Decision Tree < Lasso Regression < Ridge Regression < Linear Regression < KNN (hypertuned with suitable K-value) < Decision Tree hypertuned with Grid Search CV < Random Forest < Random Forest hypertuned with Randomized Search CV.

7. Conclusion and Recommendations

The integration of AI and ML presents a monumental opportunity to revolutionize how India manages its air quality. By moving towards a predictive, precise, and proactive system, the country can more effectively protect public health, optimize resource allocation, and accelerate its journey towards cleaner air.

Recommendations:

1. **Invest in a National Data Infrastructure:** Create a unified, open-data platform aggregating all relevant environmental, meteorological, and urban data.
2. **Fund R&D and Pilot Projects:** Encourage public-private partnerships between government bodies (CPCB, MoEFCC), academic institutions (IITs), and tech companies.
3. **Build Capacity:** Train a new generation of environmental data scientists and foster collaboration between domain experts and AI specialists.
4. **Develop a Regulatory Framework for AI Models:** Establish standards for model validation, accuracy, and transparency to ensure reliable deployment in policy.
5. **Integrate AI Outputs into Governance:** Mandate the use of AI-based forecasting and source apportionment in the execution and dynamic adaptation of GRAP and NCAP.

8. References

1. Central Pollution Control Board (CPCB). (2022). National Air Quality Index.
2. World Health Organization (WHO). (2021). Global Air Quality Guidelines.
3. Shi, X., et al. (2015). "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting." *Advances in Neural Information Processing Systems*.
4. IBM Research. (2020). "AI for Environmental Intelligence."
5. Singh, R. P., & Kaskaoutis, D. G. (2021). "Crop Residue Burning in North-West India and Its Impact on Air Quality over Delhi." *Remote Sensing*.
6. Ministry of Environment, Forest and Climate Change (MoEFCC). (2019). National Clean Air Programme (NCAP).
7. Bai, Y., Li, Y., Wang, X., Xie, J., & Li, C. (2020). Air pollutants concentrations forecasting using LSTM and its variants. *Journal of Environmental Management*, 276, 111341.

8. Guttikunda, S. K., & Calori, G. (2017). A GIS based emissions inventory at 1 km × 1 km spatial resolution for air pollution analysis in Delhi, India. *Atmospheric Environment*, 67, 101-111.
9. Jethva, H., Torres, O., Field, R. D., & Lyapustin, A. (2018). Agricultural burning in the Indo-Gangetic Plains detected by the Ozone Monitoring Instrument (OMI). *Atmospheric Environment*, 184, 257-267.
10. Kumar, A., & Pande, B. P. (2022). Air quality prediction using LSTM-based deep learning models for urban Indian cities. *Environmental Monitoring and Assessment*, 194(4), 255.
11. Kumar, P., Morawska, L., & Birmili, W. (2020). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, 143, 105799.
12. Masih, A. (2019). Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science and Management*, 5(4), 515-534.
13. Sharma, D., Srivastava, A., & Singh, P. (2022). Source apportionment of PM2.5 in Delhi using machine learning-based receptor models. *Science of The Total Environment*, 807, 150774.
14. Shtein, A., Karnieli, A., Katra, I., & Raz, R. (2020). Estimating PM2.5 in the urban environment using satellite-based aerosol optical depth and machine learning. *Remote Sensing*, 12(6), 1024.
15. Singh, N., Banerjee, T., & Devi, M. (2021). Assessing the impact of stubble burning on the air quality of Delhi using a machine learning and receptor modeling approach. *Atmospheric Pollution Research*, 12(10), 101210.
16. World Health Organization (WHO). (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Geneva: World Health Organization.