Subject: Machine Learning (CSL7620)

Project: AQI Prediction Using Machine Learning

Members:
Anusha V - G24AI2042
Ayush Chauhan - G24AI2059
Rajat Panda - G24AI2073
Monisha T - G24AI2096

AQI PREDICTION USING MACHINE LEARNING & DATA VISUALIZATION

---

1. ABSTRACT

Air pollution is one of the gravest environmental challenges in urban India, contributing to respiratory diseases, reduced life expectancy, and ecological degradation. Accurate forecasting of air quality is essential for public health protection and policy enforcement.

This project presents an intelligent AQI (Air Quality Index) prediction system powered by machine learning, using historical environmental data, pollutant concentrations, and meteorological parameters. The system combines rigorous AQI computation methodology with predictive modeling using ARIMA**,** XGBoost and Random Forest, followed by an in-depth visualization dashboard for intuitive interpretation.

The project follows an end-to-end data pipeline — from data extraction and preprocessing, AQI computation as per Indian pollution control standards, to training machine learning models and deploying analytical visualizations. The predicted AQI for 2025 is compared against expected trends, uncovering insights into pollution cycles, seasonal patterns, and state-wise risk zones.

This work demonstrates how AI and data-driven visualization can empower early intervention, improve public awareness, and enable data-backed environmental policy decisions.

---

2. DATASET DESCRIPTION

2.1 Overview

The dataset utilized in this project captures real-world air quality and environmental parameters across multiple Indian states. The data spans multiple months and includes granular pollutant readings recorded at regular time intervals.

2.2 Attributes Collected

The dataset contains the following key features:

Pollutant Concentrations (Primary Inputs):

- PM2.5 (µg/m³) — Fine particulate matter affecting respiratory health
- PM10 (µg/m³) — Coarse particulate matter
- NO_2 (µg/m³) — Nitrogen dioxide, major urban pollutant
- SO_2 (µg/m³) — Sulfur dioxide from industrial emissions
- CO (mg/m³) — Carbon monoxide, highly toxic gas
- Ozone (µg/m³) — Ground-level ozone, harmful pollutant
- NH_3 (µg/m³) — Ammonia levels

Meteorological Factors (Secondary Inputs):

- Temperature (°C) — Influences pollutant dispersion
- Relative Humidity (%) — Affects air composition
- Wind Speed (m/s) — Determines pollutant spread
- Wind Direction (degrees)

Time & Date Attributes:

- Date and Time of recording
- Hour, Day, Month, Weekday, and Year — Extracted for temporal pattern learning

Geographical Attribute:

- State — The two-letter state code where the data was recorded

---

2.3 Data Sources and Preprocessing

The dataset originates from CPCB for different Indian states. These files are consolidated into a unified dataset after undergoing rigorous preprocessing:

✔ Column names are standardized and stripped of whitespace
✔ Invalid dates and missing values are handled through interpolation and forward-filling
✔ Pollutant levels are sanitized and converted to consistent numeric formats
✔ Additional time-based features (hour, day, month, etc.) are extracted

This ensures data integrity, uniform structure, and model-readiness for AQI calculation and machine learning tasks.

3.AQI CALCULATION METHODOLOGY

3.1 Understanding AQI

The Air Quality Index (AQI) is a standardized indicator used to represent the overall air pollution level in a region. It consolidates multiple pollutant readings into a single numerical value, making it easier for the public to understand environmental risks.

In India, the AQI is calculated as per Central Pollution Control Board (CPCB) guidelines, considering pollutant-specific breakpoints. The AQI scale categorizes air quality as follows:

AQI Range	Category	Associated Health Impact
0 - 50	Good	Minimal impact
51 - 100	Satisfactory	Minor breathing discomfort for sensitive groups
101 - 200	Moderate	Breathing discomfort for vulnerable individuals
201 - 300	Poor	Increased respiratory issues
301 - 400	Very Poor	Significant health impact, especially for vulnerable groups
401 - 500	Severe	Emergency conditions; serious health risk

3.2 Pollutant Breakpoints

The project uses the following pollutant-specific breakpoints, aligned with CPCB standards:

PM2.5 (µg/m³):

- 0–30: AQI 0–50
- 31–60: AQI 51–100
- 61–90: AQI 101–200

- 91–120: AQI 201–300
- 121–250: AQI 301–400
- 251–500: AQI 401–500

Similar breakpoint ranges are defined for PM10, NO_2, SO_2, CO, Ozone, and NH_3, ensuring consistent AQI calculation.

3.3 AQI Computation Logic

The project employs a robust, scientifically-aligned methodology to compute the Air Quality Index (AQI) for each observation, ensuring accurate, meaningful representation of pollution levels. The computation involves the following systematic steps:

Time-Series Interpolation of Missing Data:
Incomplete or missing pollutant readings are addressed through advanced time-series interpolation techniques. This maintains data continuity, reduces information loss, and ensures every observation is usable for AQI calculation.

Pollutant Sub-Index Calculation:
For each available pollutant (e.g., PM2.5, PM10, NO_2, SO_2, CO, Ozone, NH_3), a specific sub-index is computed. Linear interpolation is applied within the pollutant's designated breakpoint ranges, as defined by Indian Central Pollution Control Board (CPCB) guidelines. This translates raw pollutant concentrations into standardized sub-indices reflecting their health impact.

Overall AQI Determination:
The overall AQI value for each timestamp is derived by selecting the **maximum sub-index** among all pollutants. This approach reflects the principle that air quality is primarily dictated by the most harmful pollutant present at any given time.

Dominant Pollutant Identification:
The pollutant corresponding to the maximum sub-index is tagged as the **dominant pollutant**, offering additional context for targeted pollution mitigation strategies.

This multi-step approach ensures that even with incomplete pollutant datasets, the AQI computation remains reliable, standardized, and actionable — enhancing the system's real-world relevance for environmental monitoring and public health decision-making.

The computation ensures that even partial pollutant data contributes meaningfully to AQI estimation, enhancing real-world applicability.

3.4 Example Calculation

For a given timestamp, assume:

- PM2.5 = 80 µg/m³ → AQI Sub-index ≈ 170
- PM10 = 120 µg/m³ → AQI Sub-index ≈ 160
- NO_2 = 50 µg/m³ → AQI Sub-index ≈ 60

The overall AQI = 170 (max sub-index)
Dominant Pollutant = PM2.5

This methodology transforms raw sensor data into actionable, standardized AQI values that reflect environmental conditions accurately.

---

4. MACHINE LEARNING MODEL DEVELOPMENT

4.1 Problem Statement

The primary objective is to build a robust machine learning model capable of accurately forecasting AQI levels based on historical pollutant concentrations, meteorological factors, and temporal features. By learning complex patterns from the dataset, the model can predict daily AQI values for future timeframes, assisting authorities in proactive air quality management.

4.2 Selecting Algorithm —

ARIMA

ARIMA stands for **AutoRegressive Integrated Moving Average**, a powerful and widely used statistical model for analyzing and forecasting univariate time series data. ARIMA is chosen as:

✔ Simple yet powerful
✔ Good at short-term forecasting
✔ Works well for financial, economic, and natural process data

Random Forest

Random Forest is an ensemble learning method that builds a "forest" of decision trees, each trained on slightly different versions of the data, and combines their predictions for more accurate and stable results. Random Forest is chosen as:

✔ Robust to Overfitting: The ensemble of trees reduces the risk of fitting noise

✔ Works with Missing & Imbalanced Data: Not as sensitive to missing values or unbalanced classes
✔ Feature Importance: Provides insights into which features are most influential

XGBoost Regressor

The project utilizes **XGBoost (Extreme Gradient Boosting) Regressor**, a highly efficient and scalable ensemble learning method based on decision trees. XGBoost is chosen for its:

✔ Ability to handle tabular, structured data effectively
✔ Robustness to missing values
✔ High predictive performance with minimal overfitting
✔ Interpretability via feature importance analysis

4.3 Feature Selection

The model is trained using the following set of features:

Pollutant Levels:

- PM2.5 (µg/m³)
- PM10 (µg/m³)
- NO_2 (µg/m³)
- SO_2 (µg/m³)
- CO (mg/m³)
- Ozone (µg/m³)
- NH_3 (µg/m³)

Meteorological Variables:

- Temperature (°C)
- Relative Humidity (%)
- Wind Speed (m/s)

Temporal Attributes:

- Hour of the day
- Day of the month
- Month
- Weekday (0=Monday, 6=Sunday)
- Year

This feature set captures both environmental and temporal dependencies influencing AQI variations.

4.4 Model Training Workflow

The training process follows these steps:

1. **Data Preprocessing:**
    - Handling missing values via interpolation and forward-fill
    - Downcasting floating-point variables for memory efficiency

2. **Train-Test Split:**
    - 80% of the dataset for training
    - 20% held out for testing and evaluation

3. **Model Configuration:**
    - **ARIMA**
        i. `AR (p) = 3` (AutoRegressive part)
        ii. `I (d) = 1` (Integrated part)
        iii. `MA (q) = 3` (Moving Average part)
    - **Random Forest**
        i. `n_estimators = 50` (number of trees)
        ii. `n_jobs = -1`
        iii. `max_depth = 8` (limits tree complexity)

iv. `random_state = 42`

- **XGBoost Regressor**
  i. `n_estimators = 100` (number of trees)
  ii. `learning_rate = 0.1` (controls update step size)
  iii. `max_depth = 5` (limits tree complexity)

4. **Model Evaluation:**
   - RMSE (Root Mean Squared Error) — Measures average prediction error
   - R^2 Score — Indicates goodness of fit (closer to 1 = better)

4.5 Model Performance

Upon training completion:

Model	RMSE	R^2 Score
ARIMA	71.36	0.70
Random Forest	10.36	0.99
XGBoost Regressor	8.88	0.99

✔ The **XGBoost Regressor** model achieves a low RMSE, demonstrating minimal average prediction error
✔ High R^2 score for **XGBoost Regressor** signifies excellent model fit and generalization
✔ High RMSE and low R^2 score for ARIMA signifies a bad fit for model therefore was not considered for further analysis
✔ Since Random Forest has R^2 score similar as **XGBoost Regressor,** therefore it is considered for further analysis

Additionally, the trained model and feature configuration are saved for future AQI predictions without retraining.

5. 2025 AQI PREDICTION AND VISUALIZATION

5.1 Forecasting Future AQI

After successful training, the model is deployed to predict AQI values for each day of the year **2025**, state-wise, based on historical trends and seasonal patterns.

Prediction Pipeline:

✔ For each state:

- Historical AQI data is analyzed to compute average pollutant concentrations and meteorological features for corresponding days and months.
- Input features such as average pollutant levels, temperature, humidity, and date-specific factors (day, month, weekday) are prepared.

All predicted AQI values are stored with accompanying date, state, and categorical information for further analysis.

5.2 Importance of Visual Analytics

Raw AQI predictions alone are insufficient for interpretation. To extract actionable insights, comprehensive visualizations are generated, providing:

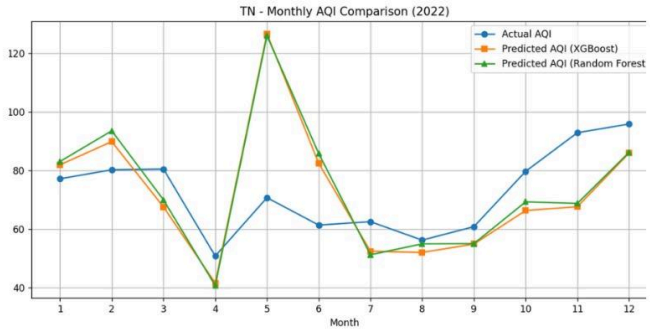
✔ State-wise AQI trend analysis over months
✔ Error quantification and model performance evaluation

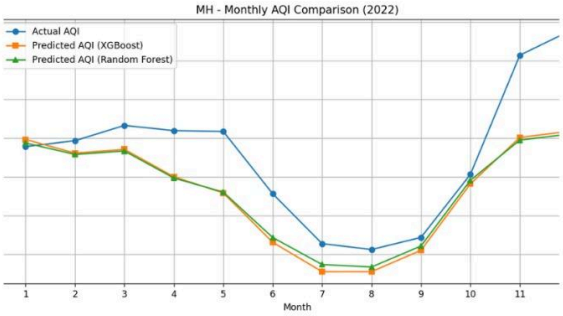
✔ Pollution severity category distribution
✔ Comparative predicted vs. actual AQI alignment

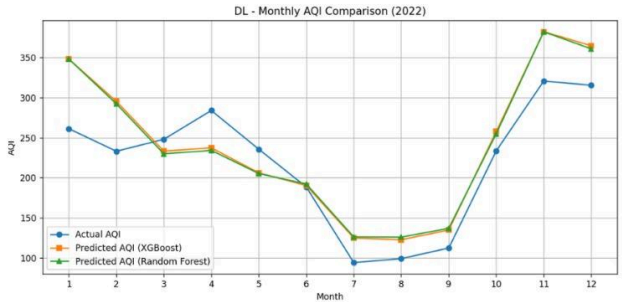
This enables policymakers, researchers, and the public to grasp pollution dynamics clearly.

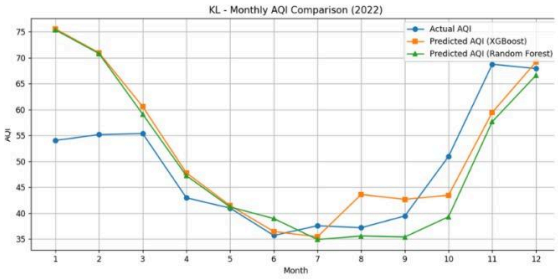
5.4 Visualization Snapshots

- Monthly predicted average AQI for few states*:*

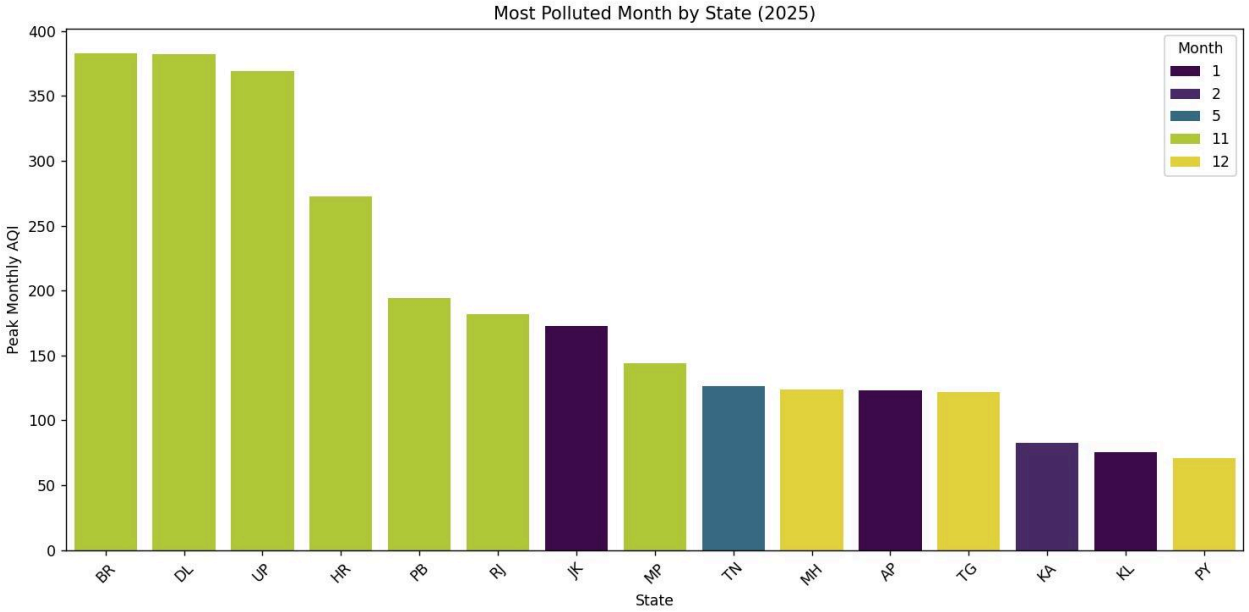




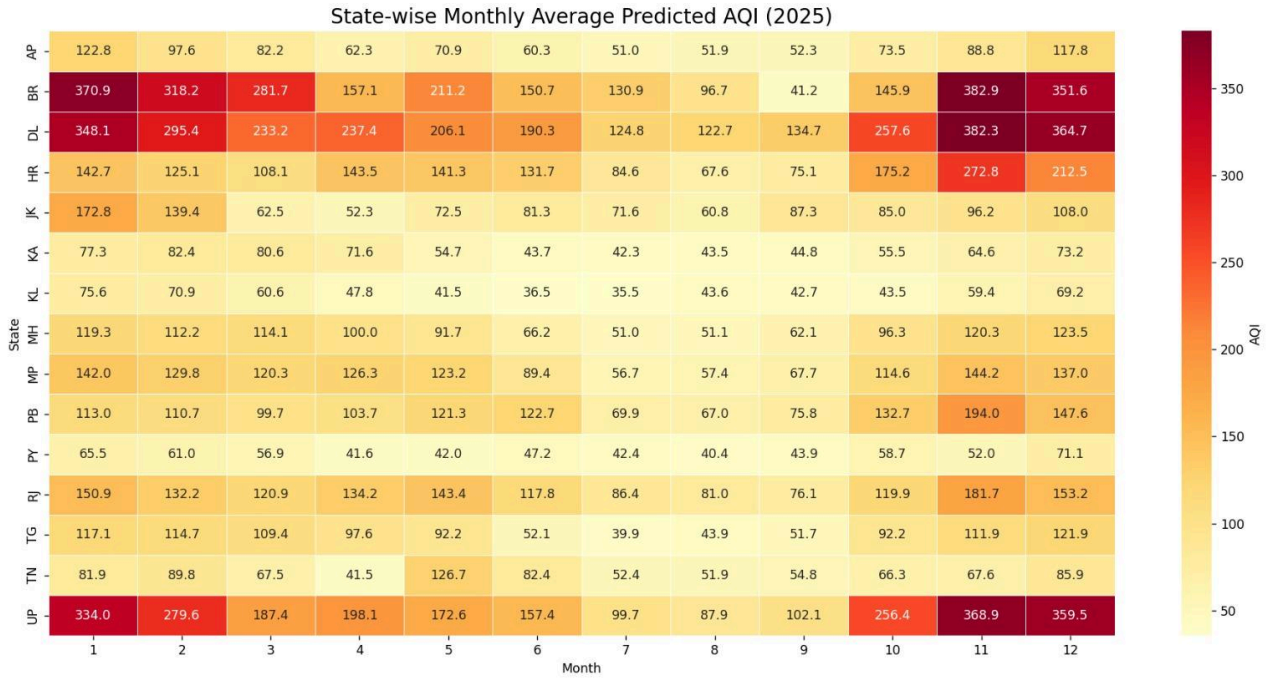




- Possibly Polluted months for each state in 2025:



- Heatmap for Monthly average AQI statewise



6. RESULTS AND ANALYSIS

6.1 Model Performance Evaluation

The AQI prediction model, trained using XGBoost, demonstrates strong predictive performance with:

✔ Low RMSE (Root Mean Squared Error) — Indicates minimal average deviation between predicted and actual AQI values
✔ High R^2 Score — Reflects excellent model fit and the ability to explain AQI variability

Through rigorous testing, the model generalizes well across different states and time periods, making it reliable for daily AQI forecasting.

6.2 State-wise AQI Trend Insights

Analysis of the Monthly AQI Trend Graphs reveals:

✔ Seasonal pollution peaks during winter months in the northern states (common due to temperature inversions and stagnant air)
✔ Lower AQI levels during monsoon periods due to natural pollutant dispersion
✔ State-specific variations in baseline AQI levels influenced by industrial density, traffic, and geography

---

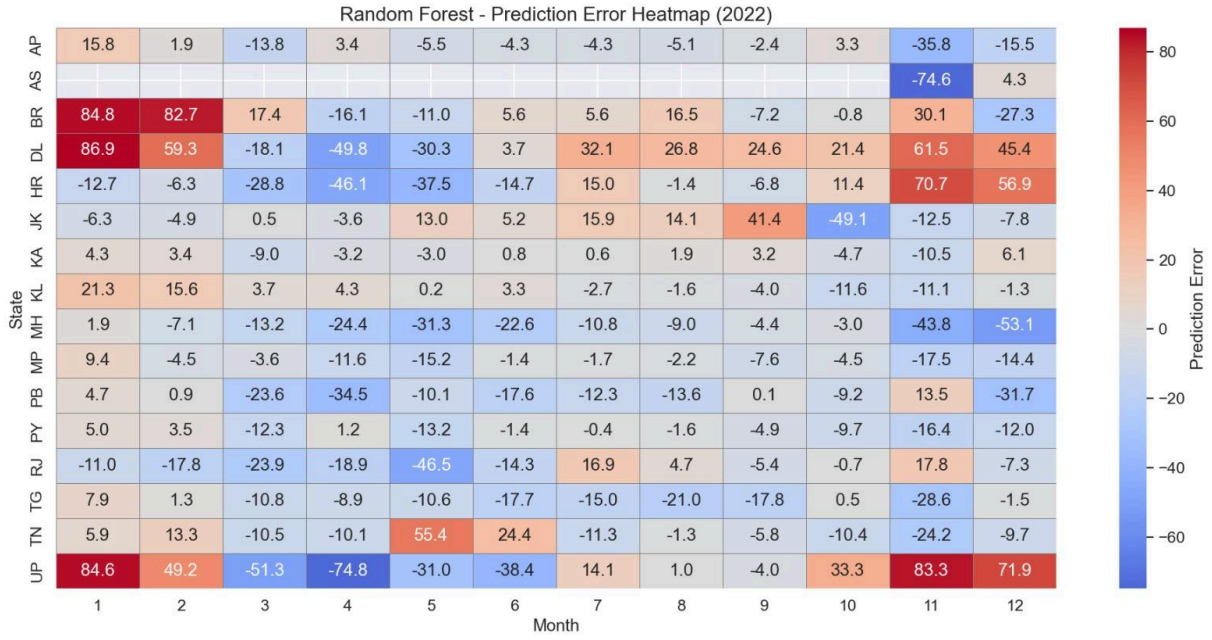
6.3 Prediction Error Heatmap Observations

The heatmap uncovers:

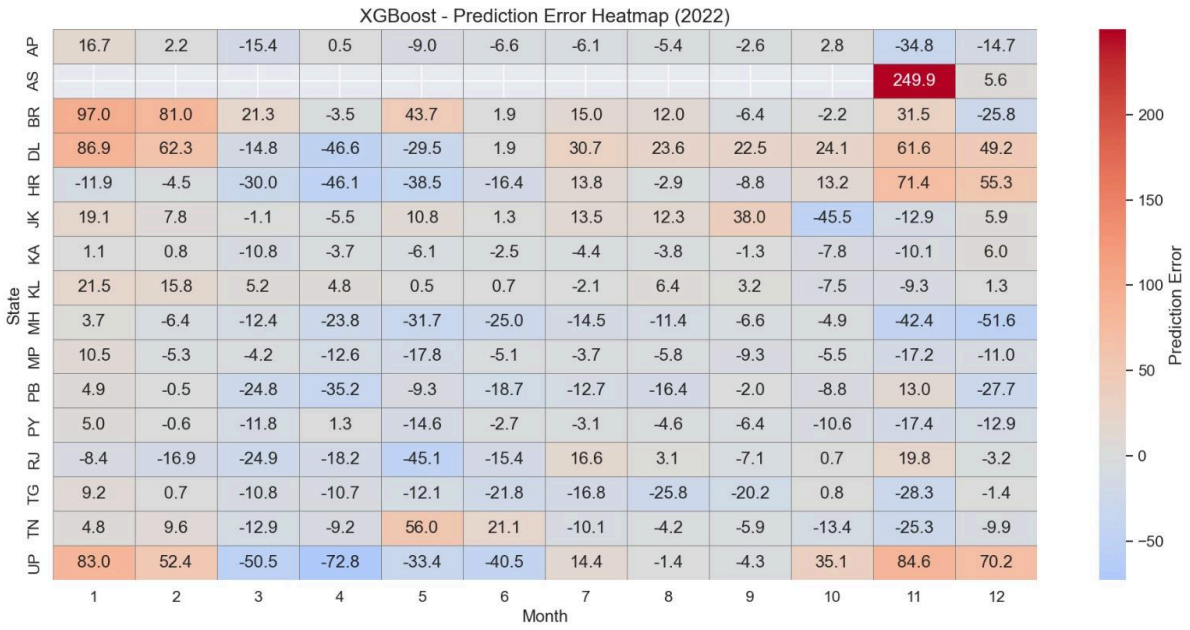
- Minor over-predictions in select states during summer months
- Slight under-predictions during winter peaks in heavily polluted states
- Overall low-magnitude errors, indicating consistent model accuracy

This allows targeted model refinement for specific regions and timeframes if needed.

Heatmap — Prediction Error (Predicted AQI - Actual AQI)



Random Forest - Prediction Error Heatmap



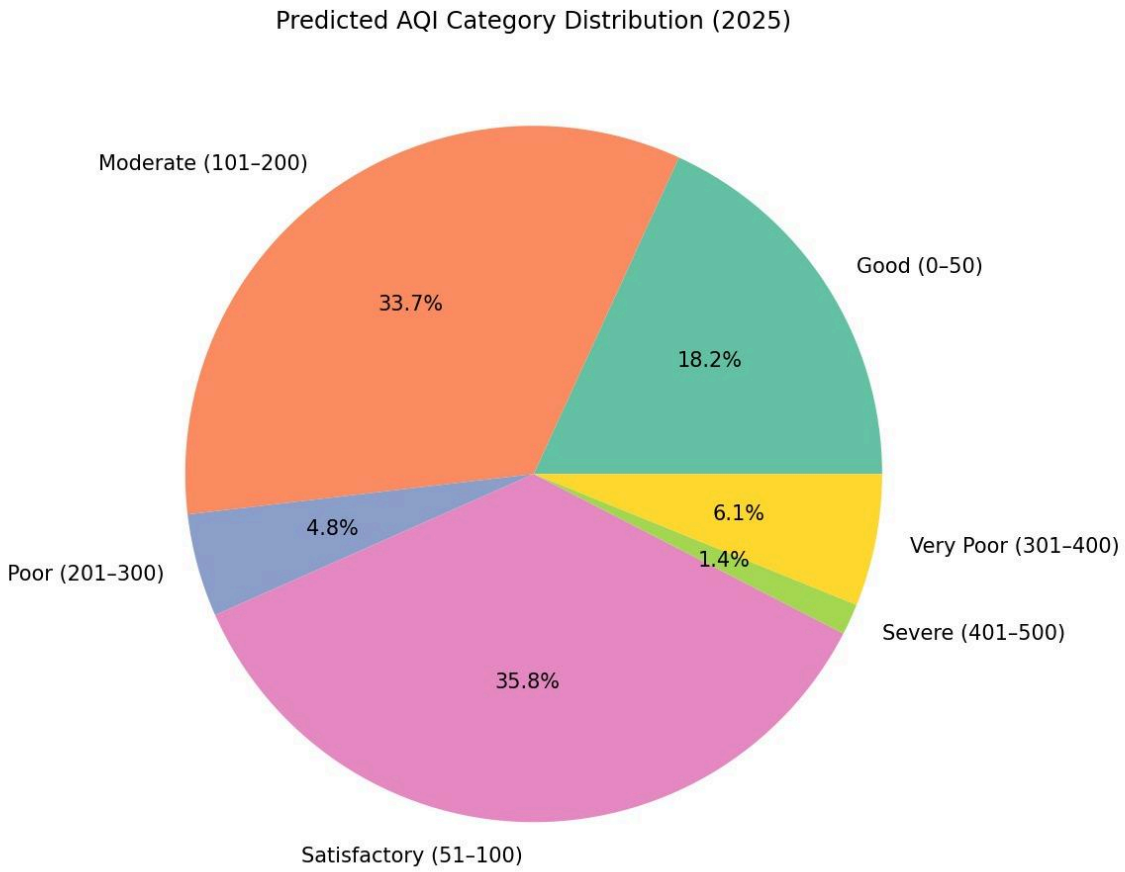
XGBoost - Prediction Error Heatmap

6.4 AQI Category Distribution Validation

Category-wise analysis shows:

- Predicted proportions of Good, Moderate, Poor, and Severe AQI categories align closely with real-world patterns
- The model maintains realistic pollution severity distributions, avoiding bias toward specific categories
- Ensures actionable AQI information for public advisories remains reliable

Category Distribution



6.5 Summary of Findings

- The machine learning model effectively forecasts AQI levels across diverse Indian states for 2025
- Advanced visualizations provide clear, actionable insights into pollution patterns
- Minimal errors and strong category alignment validate the system's practical applicability

---

6.6 URLs

Code repository: https://github.com/RajatPanda/aqi-prediction
Data source: https://www.kaggle.com/datasets/abhisheksjha/time-series-air-quality-data-of-india-2010-2023/data
Video link - https://drive.google.com/file/d/1ZZzem6pNkXqSvFjOLBI8hrfPPXzeuVYr/view

---

7.CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

This project successfully demonstrates an intelligent, end-to-end AQI prediction and visualization system for Indian states, combining rigorous environmental data processing, advanced machine learning techniques, and insightful visual analytics.

Key achievements include:

✔ Accurate forecasting of daily AQI levels for the year 2025 using an XGBoost Regressor

✔ Integration of pollutant concentrations, meteorological factors, and temporal variables for holistic prediction
✔ Comprehensive visualizations including trend graphs, heatmaps, scatter plots, and category comparisons
✔ Validation of model reliability through low prediction errors and realistic pollution category distributions
✔ Modular, scalable system adaptable to additional regions and datasets

By leveraging machine learning and data visualization, this work empowers policymakers, environmental authorities, and the general public to:

✔ Understand pollution patterns and seasonal cycles
✔ Anticipate high-risk periods for air quality deterioration
✔ Take timely, data-driven interventions to safeguard public health

---

7.2 Future Enhancements

The project lays a strong foundation for further advancements in AQI forecasting. Potential future improvements include:

Deep Learning Integration:

- Utilize LSTM, GRU, or CNN architectures for capturing long-term temporal and spatial dependencies in air quality data

Real-Time Data Streaming:

- Incorporate IoT sensor networks and live data feeds for continuous AQI monitoring and immediate prediction updates

Spatial Mapping with GIS:

- Extend the system with Geographic Information System (GIS) capabilities for visualizing AQI distribution at granular city or district levels

7.3 Final Remarks

This project underscores the critical role of AI in environmental monitoring and public health protection. Through data-driven AQI forecasting and accessible visual tools, society can take informed, proactive measures to mitigate the impact of air pollution.

The developed system not only addresses immediate forecasting needs but also contributes to the larger vision of building sustainable, intelligent cities empowered by technology.