Model Name: Random Forest Air Quality Forecasting Model
Version: 1.0
Model Type: Regression
Developers: ENG M 680 Project Team 1
Release Date: December 2025

Intended Use

This model predicts the next hour(1 hour ahead) air quality using historical pollutant measurements, meteorological variables, and station metadata. It supports environmental monitoring teams and analysts in understanding short-term air quality fluctuations.

Intended Users:
Environmental analysts, public-health researchers, and data science practitioners.

Out-of-Scope Use:
Long-term forecasting, medical diagnosis, real-time safety-critical decision automation, and regions not represented in the training dataset.

Model & Data Description

Dataset:
The final dataset contained **1,461,699 rows** after cleaning, spanning multiple cities with varying pollution conditions.

Features Used:

- Meteorological: temperature, humidity, windspeed, dew point, pressure
- Pollutants: PM10, NO2, SO2, CO, O3
- Engineered temporal features: hour of day, month
- Lag features: previous-hour PM2.5
- Station metadata: elevation, region type, etc.

Target Variable: PM2.5 concentration forecast one hour into the future (µg/m^3)

Model Architecture:
Random Forest Regressor

- 300 trees
- Max depth: unrestricted
- min_samples_split = 2

- random_state = 42

Training & Evaluation

Training Setup:
60/20/20 train-validation-test split, 5-fold cross-validation, trained in Python (scikit-learn) on JupyterHub.

Evaluation Metrics (Test Set):

- **MAE:** 21.11
- **RMSE:** 31.87
- **Baseline (Persistence):** MAE **32.54**, RMSE **46.01**

The Random Forest model **significantly outperformed** the persistence baseline, reducing forecasting error by over **10 units MAE** and **14 units RMSE**.

Ethical & Practical Considerations

Fairness:
Air-quality datasets may contain spatial biases since some regions have denser monitoring networks. Temporal irregularities and missing values were corrected, but uneven sampling remains a risk.

Privacy:
No personal data involved.

Security:
The model is not yet deployed in a public-facing API. Vulnerabilities like model extraction or tampering were not assessed.

Limitations & Recommendations

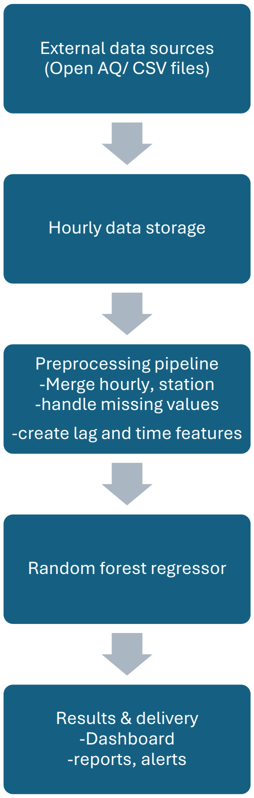
Limitations:

- Performance declines under extreme pollution spikes not seen in training data.
- Short-term forecasting only — not suited for multiday predictions.
- Weather-driven variability limits generalization across unseen geographies.

Recommendations:

- Retrain periodically with fresh data to capture seasonal shifts.
- Consider deploying alongside drift monitoring.
- Use SHAP or feature-importance tools for better interpretability.

Architecture diagram:


External data sources
(Open AQ/ CSV files)
Hourly data storage
Preprocessing pipeline
-Merge hourly, station
-handle missing values
-create lag and time features
Random forest regressor
Results & delivery
-Dashboard
-reports, alerts