

■■ Islamabad Air Quality Prediction System

This report provides a complete overview of the Islamabad Air Quality Prediction (AQI) System, developed using Python, Streamlit, and Machine Learning. The system automatically fetches real-time AQI data, performs feature engineering, trains a Random Forest model, and visualizes both real-time and forecasted AQI trends through an interactive dashboard.

Prepared by: **Abdullah Ejaz Khan**

Location: Islamabad, Pakistan

1. Project Overview

The Islamabad AQI Predictor is an end-to-end data science application designed to forecast the Air Quality Index for the next 72 hours (3 days). It uses real-time air quality data from OpenWeatherMap APIs and employs a Machine Learning model to predict future air pollution trends.

Core objectives:

- 1 Automate real-time data fetching for Islamabad using OpenWeatherMap API.
- 2 Engineer relevant temporal and pollutant-based features for accurate forecasting.
- 3 Train a robust predictive model (Random Forest) and evaluate its performance.
- 4 Deploy an interactive Streamlit dashboard with AQI visualization and report generation.

2. Data Collection and Processing

Data was fetched using the OpenWeatherMap Air Pollution API, which provides pollutant concentrations such as CO, NO2, SO2, O3, and PM2.5/PM10 levels. Approximately 4500 recent data points were collected and saved in CSV format for model training.

Feature Engineering Process:

- 1 Datetime conversion and sorting.
- 2 Creation of time-based features (hour, day, month, weekday).
- 3 Lag and rolling window features for AQI trends (lag1, lag2, roll3, roll7).
- 4 Handling missing values and filtering relevant columns.

Processed data was stored in **data/processed_features.csv** and served as the input for model training.

3. Model Training and Evaluation

The system uses a **Random Forest Regressor** trained on historical AQI data. The training pipeline was implemented in **train_model.py** and includes model evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R².

Model Results:

- 1 Mean Absolute Error (MAE): 0.00
- 2 Root Mean Squared Error (RMSE): 0.04
- 3 R² Score: 1.00

The trained model and feature names were saved in the **models/** directory for deployment in the Streamlit dashboard.

4. Dashboard Design and Functionality

The Streamlit-based dashboard (**app.py**) provides a clean, responsive interface with the following features:

- 1 ■ Real-time AQI visualization for Islamabad.
- 2 ■ 3-Day (72-hour) AQI forecast with color-coded risk levels.
- 3 ■ Interactive plots using Plotly for AQI trends.
- 4 ■ PDF report generation with AQI summary and forecasted values.
- 5 ■ Optional SHAP feature importance visualization for interpretability.

Users can adjust forecast horizons, view forecast tables, and download reports directly through the dashboard.

5. Future Enhancements

- 1 Integrate LSTM or Prophet models for improved time-series accuracy.
- 2 Add multi-city comparison and map visualization.
- 3 Deploy the system on a public cloud platform such as Streamlit Cloud or AWS.

6. Conclusion

This project successfully demonstrates an end-to-end machine learning solution for predicting air quality levels in Islamabad. From automated data collection to visualization, it represents a practical, real-world environmental monitoring application. The pipeline ensures transparency, scalability, and usability for future smart city environmental initiatives.