PROJECT ABSTRACT

AQI Forecasting using Time Series Models

Air Quality Index forecasting for a chosen city from real OpenAQ station data, comparing SARIMA, Prophet and LSTM on a strict chronological split with MAE, RMSE and MAPE evaluation, delivered built-to-order with code, notebooks, dashboard, report, PPT and viva kit.

01. Title

AQI Forecasting using Time Series Models

02. Introduction / Background

Air quality advisories in Indian cities are usually reactive: the published AQI describes air people already breathed. A reliable 24-48 hour forecast would let municipal bodies, schools and citizens act before a severe episode rather than during it. AQI is a natural time-series problem because pollution has strong structure: daily traffic cycles, weekly patterns, and seasonal episodes such as winter inversions all repeat. Most student forecasting demos, however, use synthetic or toy CSVs disconnected from real monitoring infrastructure. This project closes that gap with genuine hourly station data from the OpenAQ platform and three real forecasting approaches compared honestly.

03. Problem Statement

There is no shortage of AQI dashboards showing the present; the missing piece is a student-built, methodologically sound forecaster trained on real monitoring data. Naive demos shuffle time series (leaking the future into training), quote a single flattering error number, or never show how error grows with horizon. This project addresses all three: real OpenAQ data, a strict chronological split, and per-horizon error reporting across statistical, decomposable and deep models.

04. Objectives

- Fetch hourly PM2.5, PM10, NO2 and O3 for a real monitoring station (default: a Delhi station) via the OpenAQ API v3 and compute hourly AQI with standard breakpoint tables.
- Train SARIMA (statsmodels), Prophet and an LSTM (TensorFlow/Keras) on the same chronological 80/20 split with no shuffling.
- Evaluate all models on MAE, RMSE and MAPE at 6, 24 and 48-hour horizons on the identical holdout, with residual and error-by-horizon analysis.

- Serve a 48-hour forecast with AQI category bands on a Flask dashboard and deliver the complete student kit: code, notebooks, report, PPT and viva Q&A.

05. Proposed Methodology

Raw hourly readings are resampled to a regular index; short gaps are interpolated by documented rules and long gaps excluded. Hourly AQI is computed from pollutant concentrations via the standard breakpoint tables with the dominant pollutant recorded. The series is split chronologically 80/20. SARIMA orders are selected by AIC; Prophet uses daily/weekly seasonality with high-pollution event regressors; the LSTM trains on sliding 72-hour windows to predict 24 hours ahead. All three forecast the holdout at three horizons and are scored identically; the best-validation model drives the dashboard forecast.

06. System Architecture / Working

The fetch notebook pulls station data from OpenAQ and stores it as CSV. Preprocessing resamples, flags gaps and computes AQI. The training notebooks fit SARIMA, Prophet and the LSTM on the chronological training portion. The evaluation notebook scores all models on the holdout at 6/24/48-hour horizons and produces the metric tables and diagnostic plots for the report. The Flask dashboard loads the selected model, generates the 48-hour forecast, and renders it with category bands, dominant-pollutant info and per-model validation metric cards.

07. Hardware / Software Requirements

- Python 3.10 with pandas, NumPy, scikit-learn, Matplotlib
- statsmodels (SARIMA), Prophet, TensorFlow/Keras (LSTM)
- Jupyter notebooks for fetch, EDA, training and evaluation
- Flask dashboard application
- Free OpenAQ API key (or bulk S3 archive download); internet for data fetch
- Standard laptop/desktop; no GPU required (CPU training is sufficient)

08. Expected Outcomes

A documented pipeline from real station data to an honest three-model comparison and a working 48-hour forecast dashboard. No accuracy is pre-claimed: MAE, RMSE and MAPE are computed by the evaluation notebook on the buyer's holdout during the build, and the report presents them per horizon including the expected degradation at 48 hours. The error-by-horizon curves, residual diagnostics and missing-data audit give the student defensible, viva-ready material.

09. Applications

- Prototype early-warning display for schools, RWAs and citizen groups (advisory scope, not regulatory)
- Teaching material for time-series coursework: stationarity, seasonality, ARIMA orders, leakage
- Reference pipeline reusable for any OpenAQ station worldwide
- Foundation for multivariate extensions (weather, traffic regressors)

10. Future Scope

- Adding meteorological regressors (wind, boundary-layer height) to Prophet/LSTM
- Multi-station spatial models and city-average forecasting
- Probabilistic forecasts with prediction intervals
- Streaming deployment with scheduled retraining
- CPCB breakpoint variant alongside the US EPA computation