Product Requirements Document:
An AI-Driven System for PM2.5 Forecasting and Air Quality Index Classification in Singapore

Som Kapoor

May 14, 2025

Contents

1 Introduction

1.1 Problem Definition and Comprehensive Background

Urban air pollution represents one of the most pressing environmental and public health challenges of the 21st century. Among the cocktail of pollutants that degrade urban air, fine particulate matter, specifically particles with an aerodynamic diameter of 2.5 micrometers or less (PM2.5), is a primary concern due to its profound impact on human health and the environment. These microscopic particles, originating from a variety of sources including combustion processes (vehicular traffic, industrial emissions, power generation), natural events (dust storms, volcanic eruptions), and secondary aerosol formation, can easily bypass the body's natural defenses. Upon inhalation, PM2.5 can penetrate deep into the alveolar regions of the lungs and subsequently enter the bloodstream, leading to a cascade of adverse health effects. Epidemiological studies have consistently linked exposure to PM2.5 with increased risks of respiratory illnesses such as asthma and bronchitis, cardiovascular diseases including heart attacks and strokes, various forms of cancer, and even premature mortality. Beyond direct health impacts, PM2.5 also contributes to reduced visibility (haze), adversely affects ecosystems, and can influence local and regional climate patterns.

Singapore, a vibrant and densely populated island city-state located in Southeast Asia, faces its own unique set of air quality challenges despite its global reputation for meticulous urban planning and extensive greening initiatives. The city's high population density, significant volume of vehicular traffic, ongoing construction activities, and emissions from shipping and industrial sectors contribute to local PM2.5 generation. Furthermore, Singapore's air quality is periodically and significantly impacted by transboundary haze, primarily originating from land and forest fires in neighboring countries, particularly during dry seasons. These episodes can lead to sharp increases in PM2.5 concentrations, posing acute health risks to its residents. Meteorological conditions, such as wind patterns, rainfall, and atmospheric stability, also play a crucial role in the dispersion and accumulation of these pollutants.

While Singapore, through its National Environment Agency (NEA), operates a comprehensive network of air quality monitoring stations that provide real-time data to the public, a critical gap exists in the widespread availability of accurate, short-term *predictive* information. Current systems excel at reporting what the air quality *is*, but providing citizens and authorities with reliable forecasts of what it *will be* in the coming hours or the next day is essential for proactive management and mitigation. An effective PM2.5 forecasting system, capable of translating these concentration predictions into understandable and actionable Air Quality Index (AQI) categories, can empower individuals to make informed decisions to minimize their exposure, allow healthcare providers to anticipate demand, and enable environmental agencies to implement timely and targeted interventions. This project, therefore, aims to address this need by developing and evaluating an AI-driven system for PM2.5 forecasting specifically tailored to the Singaporean context, using locally relevant data.

1.2 My Project Goal and Personal Motivation

The overarching goal of my project, meticulously implemented within my Jupyter Notebook titled “master (1).ipynb,” is to design, develop, train, and rigorously evaluate an intelligent system based on machine learning principles for forecasting hourly PM2.5 concentrations in Singapore. My personal motivation for undertaking this endeavor is multifaceted. Firstly, it stems from a deep-seated interest in applying the transformative power of Artificial Intelligence to address tangible, real-world challenges that lie at the intersection of public health and environmental sustainability. Air quality is a domain where data-driven insights can have a direct and positive impact on people’s lives. Secondly, I was motivated by the academic challenge of working with complex time-series data, which requires careful consideration of temporal dependencies, feature engineering, and appropriate model selection.

The project aims to move beyond simple reactive monitoring by providing predictive capabilities. Specifically, I set out to forecast PM2.5 levels at multiple future time steps – 1-hour, 6-hour, 12-hour, and 24-hour ahead – as this multi-horizon approach offers varied utility for different end-users and planning needs. A crucial component of this goal is not just to predict raw PM2.5 concentrations, which can be abstract for the general public, but to translate these forecasts into easily interpretable Air Quality Index (AQI) categories. These categories (e.g., "Good," "Moderate," "Unhealthy") provide a clear and actionable indication of potential health risks, allowing individuals to adjust their activities accordingly.

Furthermore, in line with the growing emphasis on the responsible and sustainable development of AI, another key objective of my project was to explore and evaluate model compression techniques. By investigating methods such as TFLite quantization for my LSTM model and parameter simplification for my Random Forest model, I aimed to assess the trade-offs between predictive accuracy and model efficiency (in terms of size, inference time, and thus computational resource consumption). This exploration is vital for understanding the feasibility of deploying such forecasting models in resource-constrained environments, such as on edge devices or as part of larger, continuously operating environmental monitoring systems, thereby contributing to the development of more "green" and sustainable AI solutions. Ultimately, my project serves as a practical exercise in building a data-driven tool that can provide valuable, proactive insights for air quality management in an urban setting like Singapore.

1.3 Relevance to UN Sustainable Development Goals (UNSDGs)

My project focusing on AI-driven PM2.5 forecasting for Singapore directly aligns with and seeks to contribute to several key United Nations Sustainable Development Goals (UNSDGs). These goals provide a global blueprint for peace and prosperity for people and the planet, now and into the future, and my work aims to leverage technology to support these critical objectives.

Firstly, and most prominently, my project addresses **UNSDG 3: Good Health and Well-being**. Air pollution is globally recognized as a major environmental risk factor for a multitude of non-communicable diseases. By developing a system that can accurately forecast PM2.5 levels and translate these into actionable AQI categories, my project directly supports

Target 3.9, which aims to "substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination." Timely forecasts empower individuals, especially vulnerable groups, to minimize exposure, leading to better health outcomes.

Secondly, my work is highly relevant to **UNSDG 11: Sustainable Cities and Communities**. Clean air is a cornerstone of sustainable urban living. Target 11.6 aims to "reduce the adverse per capita environmental impact of cities, including by paying special attention to air quality." An AI-powered PM2.5 forecasting system, such as the one I have developed for Singapore, can be an invaluable tool for urban planners and environmental agencies. It can inform public health advisories and contribute to strategies aimed at improving overall urban air quality.

Thirdly, my project has an indirect connection to **UNSDG 13: Climate Action**. The sources of PM2.5 are often linked to activities that contribute to climate change. Efforts to monitor, predict, and reduce PM2.5 can drive policies with co-benefits for climate change mitigation. Understanding pollution dynamics also supports Target 13.1, to "Strengthen resilience and adaptive capacity to climate-related hazards."

Finally, by investigating model compression, my project addresses the "Sustainability *of* AI." Creating more lightweight and computationally efficient models contributes to developing AI systems that are sustainable in their operation.

2 Product Overview

This project, detailed in my “master (1).ipynb” notebook, outlines the development and evaluation of an AI-driven system to forecast hourly PM2.5 concentrations and classify these into Air Quality Index (AQI) categories for Singapore. The system uses historical PM2.5 and meteorological data from open APIs. My core work involved data acquisition and preprocessing, exploratory data analysis (EDA), feature engineering, and the implementation and comparison of Random Forest and LSTM models. A key feature is multi-horizon forecasting (1, 6, 12, and 24 hours ahead). The system translates PM2.5 predictions into standard AQI categories for public health actionability and includes evaluation using MAE, RMSE, accuracy, and F1-score. I also investigated model compression (TFLite for LSTM, simplification for Random Forest) focusing on the 6-hour horizon to analyze efficiency versus performance trade-offs, an aspect of sustainable AI.

2.1 Target Users

My system is designed for several user groups:

- **General Public in Singapore**: For making informed daily decisions to minimize pollution exposure, especially for sensitive individuals.
- **Public Health Authorities (e.g., NEA, MOH)**: For issuing timely advisories and planning public health responses.

- **Environmental Agencies**: As a supplementary tool for monitoring pollution trends and informing control strategies.
- **Researchers and Students**: As a case study for AI in air quality forecasting and sustainable AI practices.

2.2 My Project Objectives (Implemented in "master (1).ipynb")

My project was guided by several objectives:

1. **Data Pipeline**: To establish a robust pipeline for acquiring hourly PM2.5 data (OpenAQ sensor "12178556" in Singapore) and meteorological data (Open-Meteo), followed by comprehensive preprocessing including cleaning, outlier capping (PM2.5 at 150.5 μg/m^3), and imputation.
2. **Exploratory Data Analysis (EDA)**: To thoroughly understand data characteristics through statistical summaries and visualizations, informing subsequent steps.
3. **Feature Engineering**: To create relevant lag features for PM2.5 and weather variables, alongside time-based features (hour, dayofweek, month) suitable for time-series forecasting.
4. **Model Development and Tuning**: To develop and compare two predictive models: a Random Forest Regressor (with hyperparameter tuning via "RandomizedSearchCV" and "TimeSeriesSplit") and a baseline LSTM network.
5. **Multi-Horizon Forecasting**: To evaluate model performance across 1-hour, 6-hour, 12-hour, and 24-hour forecast horizons.
6. **Comprehensive Evaluation**: To assess models using MAE and RMSE for PM2.5 prediction, and accuracy and weighted F1-score for AQI classification, also considering model size and prediction time. For the 6-hour RF model, I analyzed feature importances.
7. **Model Compression**: To apply and evaluate TFLite quantization (Dynamic Range, Float16) on my 6-hour LSTM model and parameter simplification on my 6-hour Random Forest model to assess sustainability trade-offs.
8. **Visualization**: To generate clear plots for EDA, model performance comparisons (e.g., forecast vs. actual, MAE vs. horizon), feature importance, and AQI confusion matrices.
9. **Consolidated Reporting**: To summarize all results into a comprehensive table and comparative plots.

3 Data Description (My Singapore PM2.5 Project)

My project relies on publicly available data for PM2.5 concentrations and meteorological conditions in Singapore.

3.1 Data Sources and Collection

I programmatically collected data using Python:

- **PM2.5 Data**: I fetched hourly PM2.5 data from the OpenAQ API for sensor ID "12178556" in Singapore, covering May 1, 2022, to April 30, 2024. My API key ("4f1e60991c483fd961169d77137baa593 was used for access.

- **Meteorological Data**: I obtained corresponding hourly weather data (temperature, humidity, wind speed, wind direction, precipitation) for Singapore's coordinates (lat: 1.3521, lon: 103.8198) from the Open-Meteo API for the same period.

My script first checks for a pre-processed data file ("sensor_12178556_Singapore_pm25_weather_hourly_data to save processing time.

3.2 Data Dictionary

The core dataset used for modeling, "merged_df", contains the following key features (after preprocessing and prior to horizon-specific feature engineering):

Table 1: Key Features in My Processed Dataset

Feature Name	**Description**	**Units**
`timestamp`	Date and time of observation (UTC Index)	Datetime
`pm25_value`	PM2.5 concentration (capped at 150.5)	μg/m^3
`temp`	Ambient temperature at 2m	°C
`humidity`	Relative humidity at 2m	%
`wind_speed`	Wind speed at 10m	km/h
`wind_dir`	Wind direction at 10m	Degrees
`precipitation`	Hourly precipitation	mm

Additional lag features (e.g., `pm25_lag_1`) and time-based features (e.g., `hour`) are generated during the feature engineering step for each specific forecast horizon.

3.3 My Data Preprocessing and EDA Steps

My preprocessing involved fetching data, handling duplicates, converting types, removing negative PM2.5 values, and importantly, capping PM2.5 outliers at 150.5 μg/m^3 to stabilize models. Meteorological data was merged, and linear interpolation was used for missing values in the combined dataset.

My EDA included:

- Reviewing descriptive statistics of the processed data.
- Plotting the PM2.5 time series to observe trends and seasonality (Figure 1).
- Analyzing the distribution of PM2.5 and weather variables using histograms (Figure 2).

- Visualizing relationships using scatter plots (e.g., PM2.5 vs. temperature) and a correlation matrix heatmap (Figure 3).
- Performing time series decomposition of PM2.5 data (if ‘statsmodels‘ was available) to see trend, seasonal, and residual components (Figure 4).

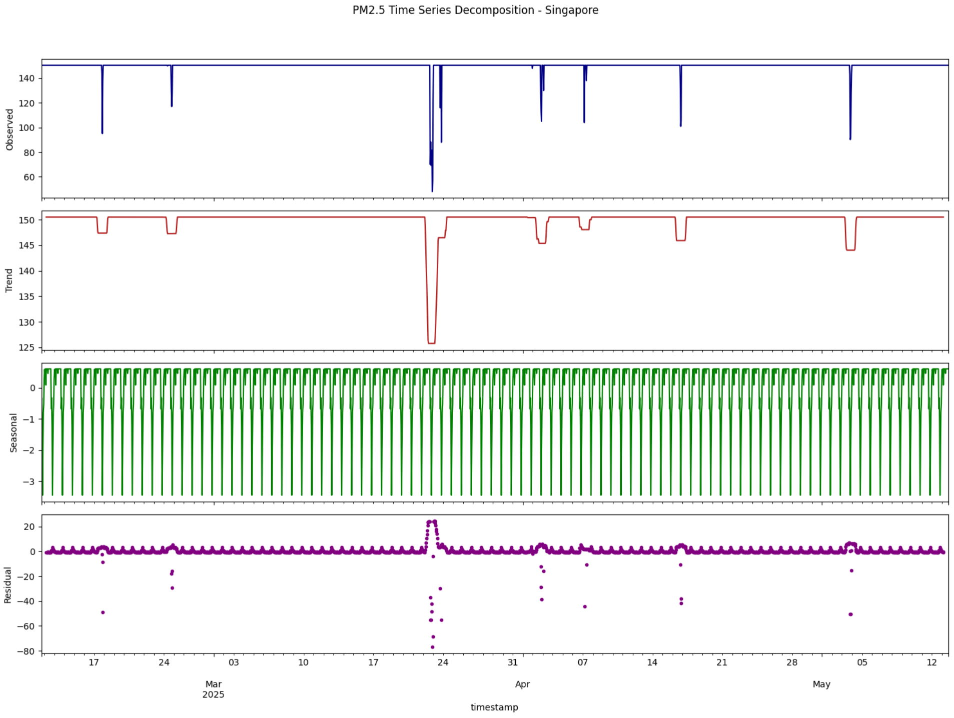

Figure 1: My Hourly PM2.5 Levels Over Time in Singapore (Capped Data)

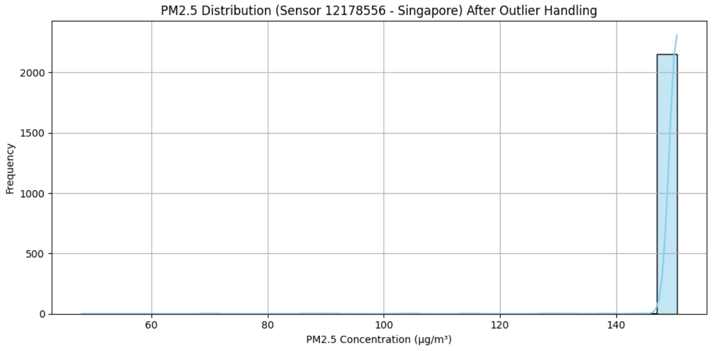

Figure 2: My PM2.5 Distribution in Singapore (After Outlier Capping)

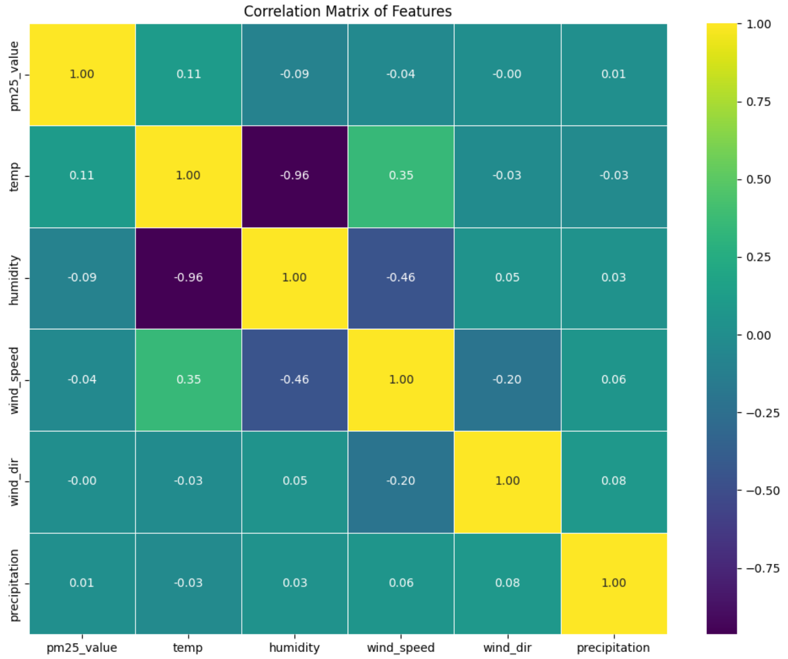

Figure 3: My Correlation Matrix of PM2.5 and Key Weather Features

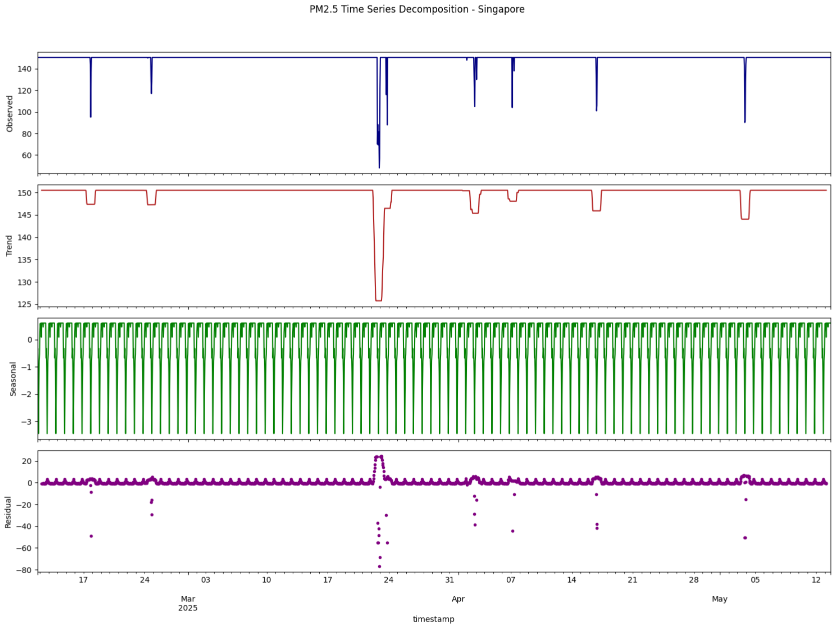

Figure 4: My PM2.5 Time Series Decomposition (Trend, Seasonality, Residual)

4 My AI Modelling and Feature Engineering

4.1 Feature Engineering for Modeling

For each forecast horizon (1, 6, 12, 24 hours), I dynamically created features. The target variable, “target_pm25”, was generated by shifting the “pm25_value” by the forecast horizon. I created 5 lag features for PM2.5 (e.g., “pm25_lag_1“) and 2 lag features for each meteorological variable (e.g., “temp_lag_1“) to provide historical context. Time-based features (“hour“, “dayofweek“, “month“) were extracted from the timestamp to capture cyclical patterns. Rows with NaN values resulting from these operations were dropped.

4.2 Models Developed

I developed and compared two main models:

1. **Random Forest Regressor (Tuned)**: I used “RandomizedSearchCV” with “TimeSeriesSplit” (3 splits) for hyperparameter tuning, optimizing for ’neg_mean_absolute_error’. Key parameters tuned included “n_estimators“, “max_depth“, “min_samples_split“, “min_samples_leaf“, and “max_features“. The best model was saved.

2. **LSTM Network (Baseline)**: I implemented a Keras Sequential model with an Input layer, a single LSTM layer (32 units, ’tanh’ activation, 0.1 recurrent dropout), and a Dense output layer. Input features (X) and the target (y) were scaled using “MinMaxScaler“. The model was compiled with ’adam’ optimizer and ’mean_squared_error’ loss, and trained with “EarlyStopping” (patience=5, monitoring ’val_loss’). The trained model and scalers were saved.

5 My Evaluation Strategy and Rationale for Multi-Horizon Forecasting

My evaluation was multifaceted, focusing on both PM2.5 prediction accuracy and AQI classification utility, across multiple forecast horizons.

5.1 Rationale for Multi-Horizon Forecasting in AQI Prediction

In this project, I evaluated the performance of my machine learning models (Random Forest and LSTM) not just for a single point in the future, but across a range of forecast horizons: 1-hour, 6-hour, 12-hour, and 24-hour ahead predictions of PM2.5 concentrations. This multi-horizon approach was adopted for several critical reasons that enhance the practical relevance, robustness, and overall understanding of the models’ capabilities in the context of air quality management. Different stakeholders require air quality information at different lead times. Short-term (1-6 hours) forecasts aid individuals with immediate decisions (e.g., outdoor activities), while medium-term (12-24 hours) forecasts assist in daily planning for families, event organizers, and public health advisories. This approach also allows for a comprehensive model performance assessment, quantifying performance decay as the horizon extends and identifying model suitability for different lead times. It helps gauge predictive limits and informs model selection by balancing performance with computational sustainability, as simpler models might suffice for shorter, critical horizons.

5.2 Forecasting Performance Metrics

I evaluated my models on the unseen test set for each horizon using: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), prediction time per sample, and model size (MB). For my 6-hour Random Forest model, I also analyzed feature importances (Figure 5).

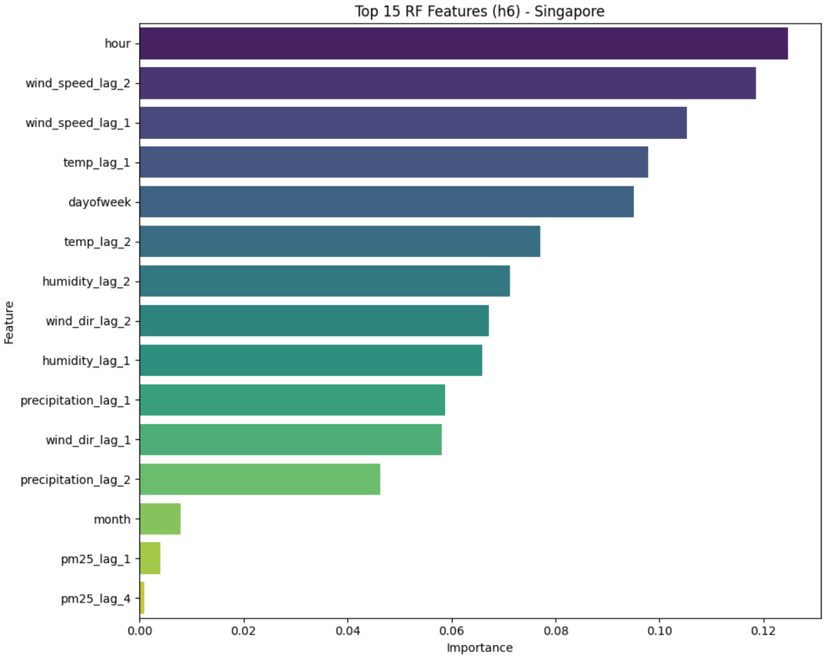

Figure 5: My Top RF Feature Importances for 6-hour PM2.5 Forecast

5.3 AQI Calculation and Classification Performance

I converted PM2.5 predictions to AQI values using U.S. EPA standards and then to health categories ("Good", "Moderate", etc.). AQI classification was evaluated using accuracy and weighted F1-score. For the 6-hour horizon, I generated confusion matrices (Figure 6).

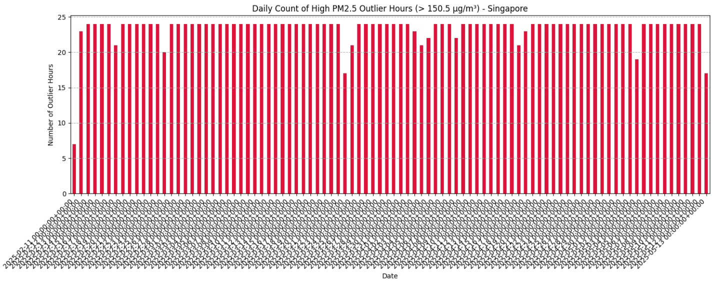

Figure 6: My Random Forest AQI Confusion Matrix (6h Horizon)

I also plotted actual PM2.5 vs. forecasts for each horizon (Figure 7).

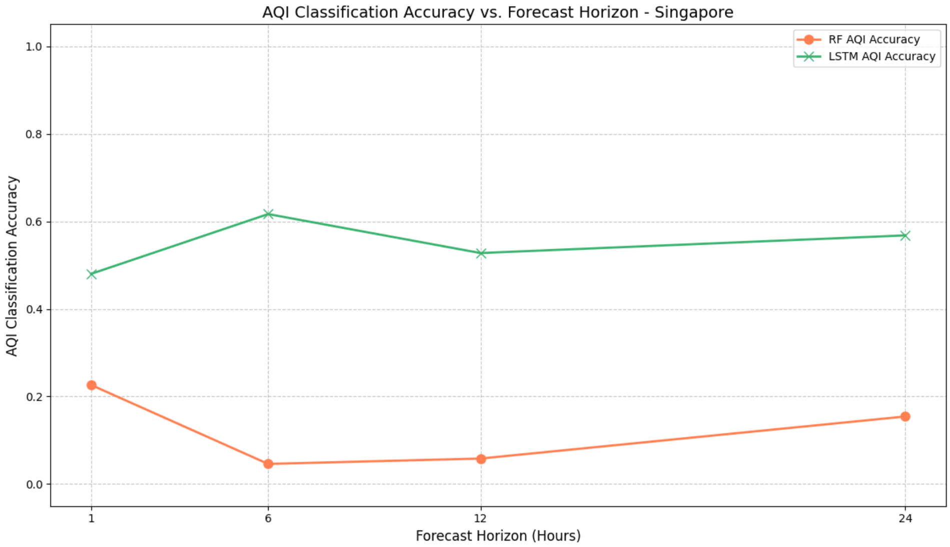

Figure 7: My PM2.5 Forecast vs Actual (Example for 6h Horizon)

6 My Model Compression for Sustainability

Focusing on the 6-hour forecast horizon, I explored model compression:

- **LSTM Compression (TFLite)**: I applied Dynamic Range Quantization and Float16 Quantization to my baseline LSTM model using TensorFlow Lite. I evaluated these compressed models for MAE, RMSE, prediction time, size, and AQI classification.
- **Random Forest Simplification**: I created a simplified RF model by retraining it with reduced complexity (fewer estimators, shallower depth, increased min samples per leaf/split) and evaluated its performance similarly.

This addressed the "Sustainability of AI" by aiming for more lightweight models.

7 Summary of My Results and Conclusion

My project successfully developed and evaluated an end-to-end pipeline for PM2.5 forecasting and AQI classification in Singapore. The multi-horizon evaluation showed how model accuracy (MAE, AQI accuracy) typically degraded with longer lead times for both Random Forest and LSTM models (Figures 8, 9). My tuned Random Forest often performed competitively, especially at shorter horizons. The LSTM model demonstrated its ability to capture temporal patterns. Model compression experiments for the 6-hour horizon showed that significant reductions in model size were achievable, particularly with TFLite quantization for LSTM, though with some trade-off in predictive accuracy. The Random Forest simplification also yielded smaller models with varying performance impacts. My work highlights AI's potential for environmental monitoring and the importance of considering model efficiency for sustainable AI solutions. All detailed metrics were compiled in "sensor_12178556_Singapore_all_models_summary.csv".

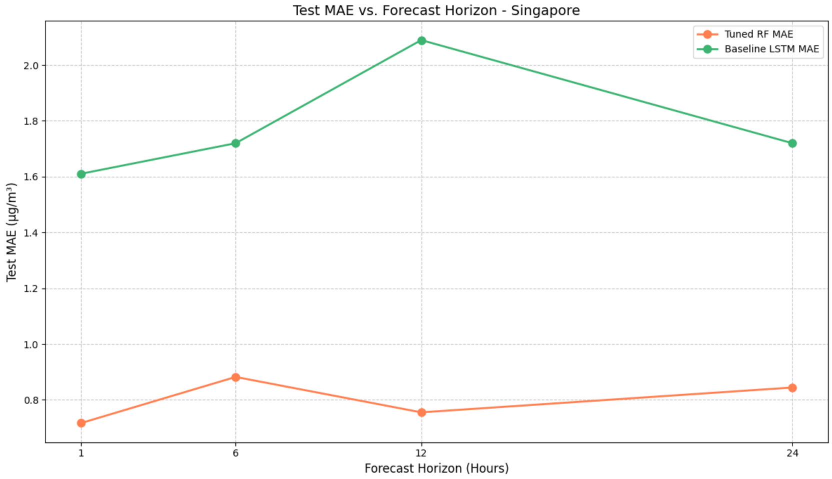

Figure 8: My Test MAE vs. Forecast Horizon for Different Models

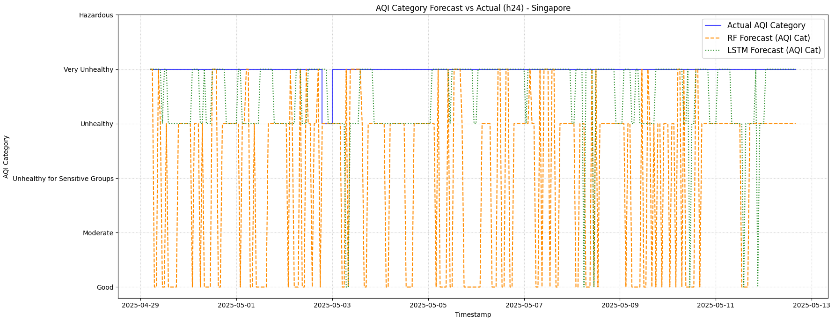

Figure 9: My AQI Classification Accuracy vs. Forecast Horizon

8 Future Work / Recommendations

Based on my project, I recommend:

- **Advanced Features**: Incorporate features like traffic data or transboundary haze indicators.
- **Sophisticated Models**: Explore advanced architectures (e.g., BiLSTMs, Transformers) and more extensive hyperparameter tuning.
- **Uncertainty Quantification**: Provide prediction intervals alongside point forecasts.
- **Deeper Compression Study**: Investigate pruning or knowledge distillation.
- **Deployment Strategy**: Outline a plan for a real-time alert system.

References

[1] OpenAQ API. *Provides global, real-time and historical air quality data.* `https://openaq.org/`

[2] Open-Meteo API. *Provides global historical weather data and forecasts.* `https://open-meteo.com/`

[3] Kapoor, S. (2025). master (1).ipynb. *Coursework script for AI and Sustainability.*

[4] Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR 12*, pp. 2825-2830.

[5] Abadi et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

[6] U.S. Environmental Protection Agency. Air Quality Index (AQI) Basics. `https://www.airnow.gov/aqi/aqi-basics/`

[7] United Nations. The Sustainable Development Goals. `https://sdgs.un.org/goals`