Air Quality Index (AQI) Prediction

Objective: Train a model to forecast AQI 3 days into the future

GitHub Repository: https://github.com/muhammadrafayasif/aqi-prediction

Submitted By: Muhammad Rafay

Dated: 9th November, 2025

Feature Selection

The following features are included in the final dataset and used for training the model. This includes relevant weather data and air pollutant data used to predict AQI.

Feature	Name (in dataset)	Unit
Temperature	temp	Celsius
Wind Speed	wind_speed	Kilometer per Hour
Wind Gusts	wind_gusts	Kilometer per Hour
Humidity	humidity_percent	Percentage
Dew Point	dew_point	Celsius
Pressure	pressure	Millibar
Cloud Cover	cloud_cover	Percentage
Visibility	visibility	Kilometers
PM_{10}	pm_10	µg/m^3
$PM_{2.5}$	pm_2_5	µg/m^3
NO_2	no_2	µg/m^3
O_3	o_3	µg/m^3
SO_2	so_2	µg/m^3
CO	co	µg/m^3
Air Quality Index	aqi	Unitless

Model Selection

Model	Selection	Reason
Random Forest Regressor	Initial model (dropped)	Heavy overfitting
Gradient Boosting	Experimented	Not generalizing
Ridge Regression	Experimented	Not generalizing
XGBoost	Final Selected	–

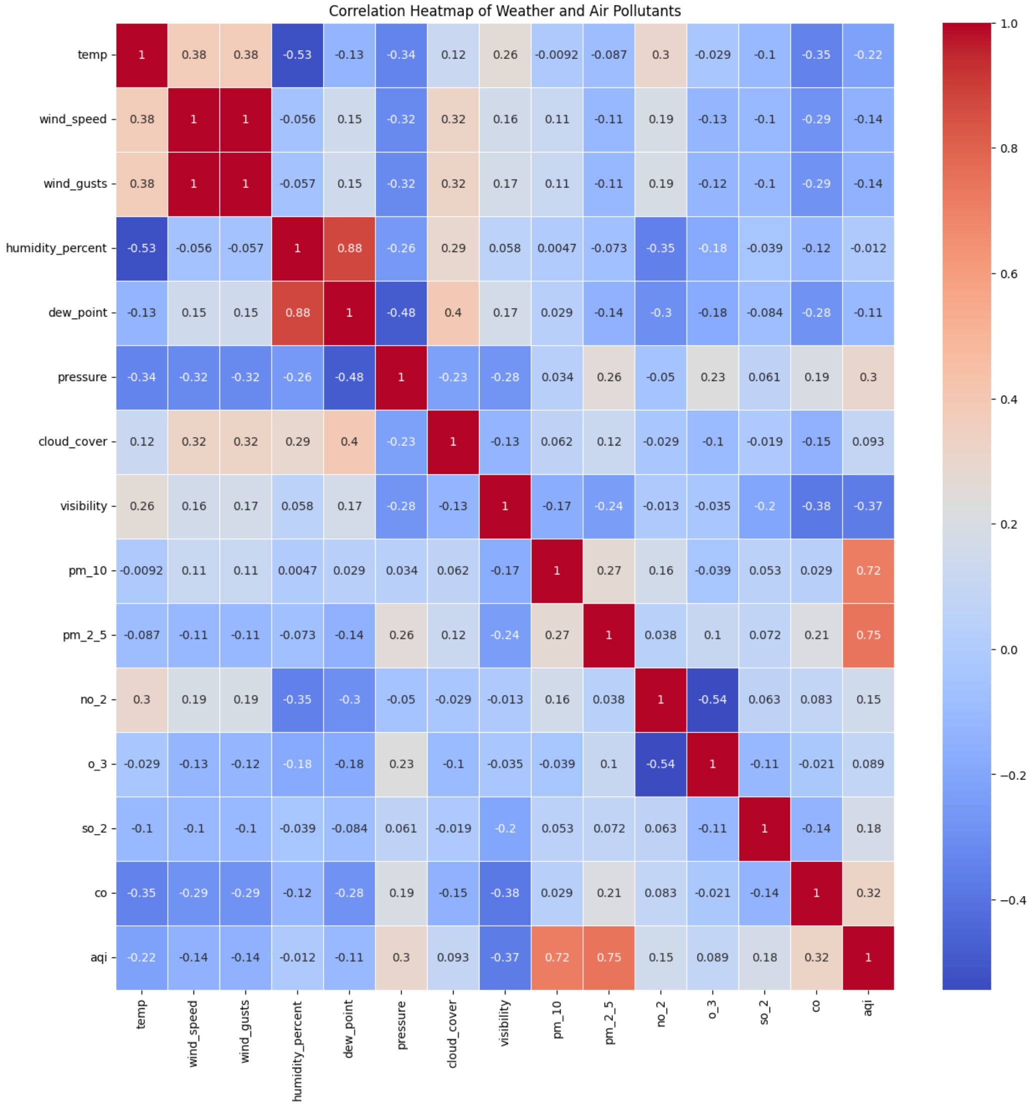
This prediction model uses XGBoost and trains separate individual models with different horizons (1, 2, 3, 6, 9, …, 72) for added accuracy.

The predictor interpolates between different horizons and joins all predictions together to create a complete 72h forecast.

Reasons for not utilizing other models

Initially, the Random Forest Regressor was chosen to be the model for forecasting, however the model had trouble learning from a small dataset (~ 1000 hours) and overfitted despite various attempts which involved pruning the tree. Eventually, the model was dropped and other options were tested.

Exploratory Data Analysis (EDA)



From the above correlation matrix, we can determine the AQI has the strongest correlations with the pollutants CO, $PM_{2.5}$ and PM_{10} and the least with O_3. Similarly, AQI is least correlated with cloud cover and humidity according to the graph.

GitHub Workflows

This repository contains two GitHub workflows and actions.

1. The data gathering pipeline runs all hours and scrapes data from AccuWeather, after successful scraping, the data is stored as a Pandas DataFrame and stored into the HopsWorks Feature Store.

2. The training pipeline gathers the data from the feature store and cleans it and handles any missing gaps (there is a 5-day gap due to an outage). It interpolates data if missing hours are found. Afterwards, it creates separate horizons (models) tailored for predictions of 1h, 2h, 3h, 6h, 9h, …, 72h and uses interpolation to estimate the future values in between each horizon.

Drawbacks and Limitations

1. The model uses interpolation in between each horizon to make a calculated estimate of the AQI value in between them. This may showcase a trend but tends to miss sudden spikes and changes as interpolations tend to be smooth.

2. The dataset is relatively small (~1000 hours). A large dataset would yield more accurate and better results from the model. Since a web scraping approach was utilized, accurate results were prioritized. As time passes and the data increases, the model will become more accurate.