PEARLS AQI PREDICTOR

PROJECT PRESENTATION

Presented By: Shafia Memon

DATA COLLECTION

- I used one year of hourly weather and pollution data for Karachi (January 2024 to January 2025), collected from the Open-Meteo API.
- The dataset included temperature, humidity, wind speed, wind direction, pressure, PM2.5, PM10, CO, NO_2, SO_2, and O_3.
- The data was cleaned and aligned by timestamp so that every hour had matching weather and pollution values.
- This prepared dataset was used to create features and train prediction models.

FEATURES

- I created time-based features such as hour, day, and month to capture daily and seasonal patterns.
- Lag features were added so the model could learn from previous pollution levels (e.g., PM2.5 from 1 hour, 3 hours, and 24 hours earlier).
- Rolling averages over 6 hours and 24 hours were used to show short-term and long-term air quality trends.
- AQI was calculated from PM2.5 and PM10 using EPA standards so that the model predicts a single air quality score.

ML MODELS

To predict AQI, I tested multiple machine learning models:

Ridge Regression

- Used as a simple baseline model.
- It is fast and easy to train, but it assumes straight-line relationships.
- Because air pollution patterns are complex and non-linear, Ridge Regression could not capture them well, so its accuracy was lower.

Random Forest Regressor

- This model uses many decision trees together.
- It is good at learning non-linear patterns and interactions between weather and pollution variables.
- It handled the data complexity better, gave high accuracy, and stable predictions.

Gradient Boosting Regressor

- Also tested to see if boosting improves accuracy.
- It performed well, but was slightly less accurate and more sensitive to noise than Random Forest.

The Random Forest model performed the best:

- R^2 Score: ~0.93
- (The model explains about 93% of the variation in AQI — good accuracy)
- MAE: ~2.33
- (On average, predictions differ from actual AQI by only around 5 points)

These results show the model predicts short-term AQI levels reliably.
Because of its high accuracy, stability, and ability to learn non-linear patterns,
Random Forest was selected as the final model for AQI prediction.

RESULT AND PIPELINE

- The final Random Forest model can reliably predict short-term AQI for Karachi.
- The model is saved in the Model Registry and can be loaded into a Streamlit dashboard.
- Pipeline steps are simple:
- Data is collected → Features are created → The model predicts AQI → The dashboard displays real-time results.
- This helps users easily check current air quality and make safer daily decisions.