AQI Prediction System – Detailed Project Report

Submitted by: Alishba Irfan , Data science intern -10 Pearls

Project Overview

The goal of this project is to build a complete end-to-end system for predicting air quality (AQI) in Karachi, combining real-time data ingestion, feature engineering, machinelearning modeling, automated pipelines, and a web-based dashboard. The system predicts both next-hour AQI and 3-day average AQI, with color-coded alerts for hazardous levels.

Key components:

Feature Pipeline – fetching and computing model-ready features.

Model Training Pipeline – training, evaluating, and storing ML models.

CI/CD Automation – scheduled updates for features and models.

Web Dashboard – interactive visualization of real-time and forecasted AQI.

Data Collection

Sources

- Air Pollutants: PM2.5, PM10, NO2, O3
- Weather Data: temperature, humidity, wind speed ,weather description
- APIs: Explored AQICN, OpenWeather, openmateo, openaq and other publicly available sources

Collection Strategy

- Data fetched hourly from APIs from weatherbit.io api and openweather api for realtime data collection
- Historical data backfilled for training ML models from weatherbit.io from January to September 2025 hourly data.
- Timestamps stored with timezone information for alignment with predictions.

Data Preprocessing

- Timestamp handling: Convert to datetime, handle missing values, forward-fill hourly gaps.
- Numeric conversion: All pollutant and weather features converted to numeric types. Weather description feature converted in numeric codes from categorical value.
- Derived features:

    - AQI change rate: difference of current aqi and previous aqi
    - PM2.5 to PM10 ratio: derived pm25_to_pm10_ratio ratio between pollutants
    - Temperature and humidity changes: derived temp_change, humidity_change
- Time-based features:
    - hour, day, month, weekday, is_weekend
- Cleaning: Remove rows with NaNs after feature computation to ensure consistent model input.

Feature store: Used hopsworks feature store for:

- Historical data stored in feature group as karachi_aqi_features which has processed data fetched from weatherbit.io api having January to September data
- For realtime data stored in feature group named karachi_realtime_features which has data fetched and computed from feature pipeline hourly_fetch.yml and used for training model in model_train pipeline
- For model storage hopsworks is used to store best_aqi_model.pkl in model registery as aqi_predictor to be used in training pipeline named model_train.yml

Feature Pipeline

- Features computed from raw and preprocessed data.
- Stored in a Feature Store hopsworks for hourly data realtime used feature group named karachi_realtime_features automated through git actions
- Feature pipeline can be scheduled hourly via CI/CD which fetches raw data from openweather and weatherbit.io and combines as realtimedata.csv and then computes it through compute.py and stores as computed_features.csv and in hopsworks feature store.

Model Training Pipeline

Training Data

- Historical features and AQI targets retrieved from the Feature Store.
- Time series split

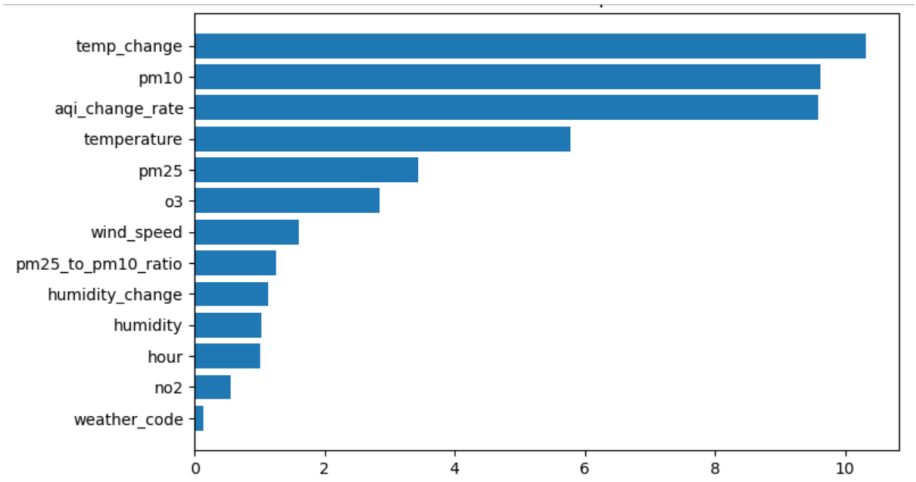
Machine Learning Models

- models: machine learning models like Linear regression, ridge regression, Random Forest, Ridge Regression, Gradient Boosting were explored on historical data.
- Evaluation metrics: Mean Absolute Error, R2, cross validation R2
  MAE (Mean Absolute Error) lower is better. On average, model's AQI predictions are off by 2.08 with XGBoost regressor
  R2 (Coefficient of Determination) closer to 1 is better. XGBoost explains 95% of the variability in AQI excellent fit.

CV_R2(Cross-Validation R2)measures how consistent the model performs on unseen data. XGBoost's 0.85 CV_R2 means it's not just memorizing training data, it generalizes well too.

- Identified which pollutants or weather features have most impact on AQI.

```
Linear Regression MAE: 7.90, R²: 0.55, CV_R2: 0.47
Ridge MAE: 7.79, R²: 0.57, CV_R2: 0.47
Random Forest MAE: 3.31, R²: 0.91, CV_R2: 0.79
XGBoost MAE: 2.08, R²: 0.95, CV_R2: 0.85
```



Model Storage

- Trained model stored in a Model Registry hopsworks with versioning.

CI/CD Pipeline

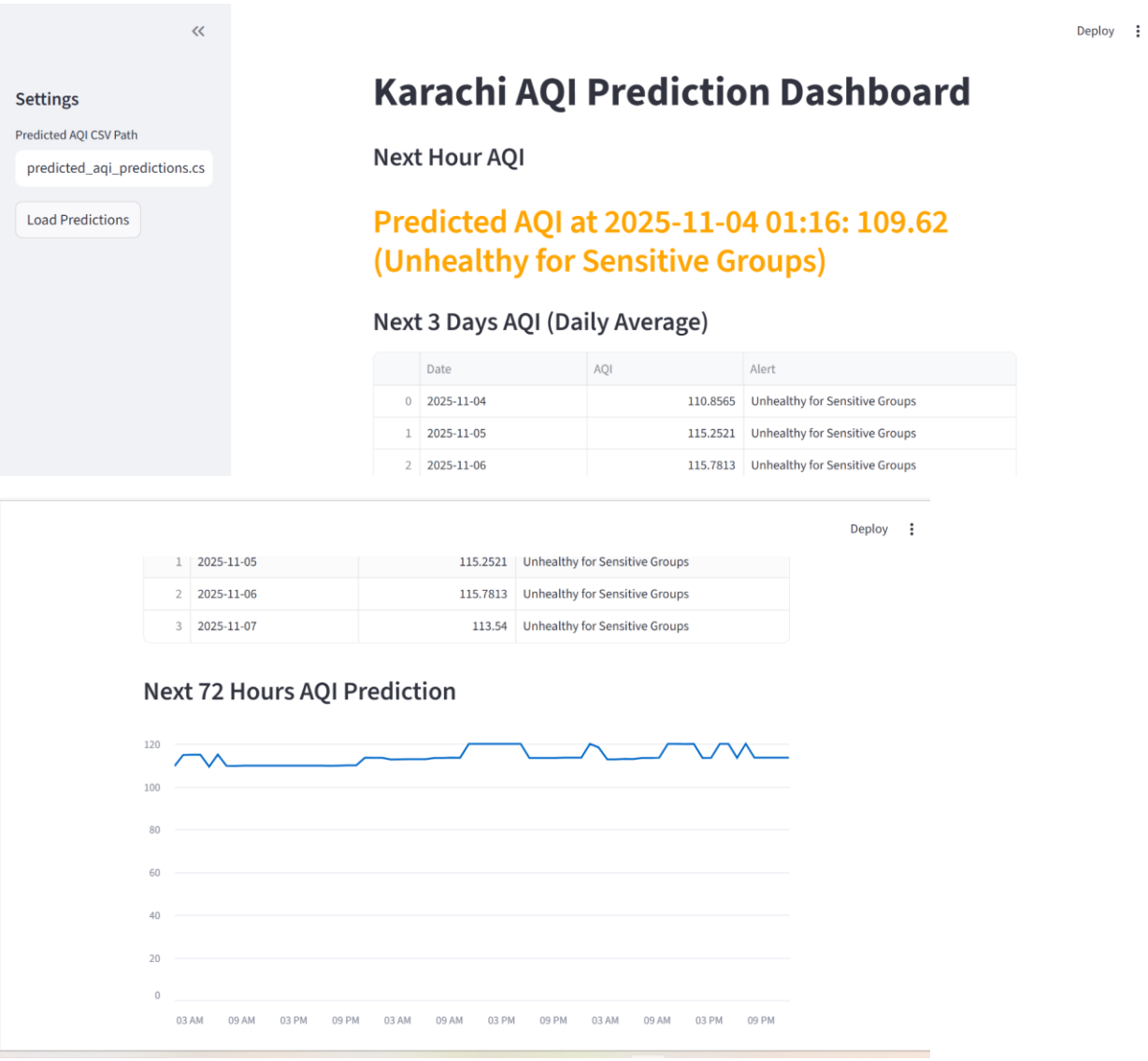
- Feature script: runs hourly to update processed features in the Feature Store.
- Training script: runs daily to retrain and store model and predict next 3 days of aqi value.
- Automation Tools**:** GitHub Actions.
- Artifacts produced:
  - Csv of latest processed features
  - Pickled trained model file and prediction csv

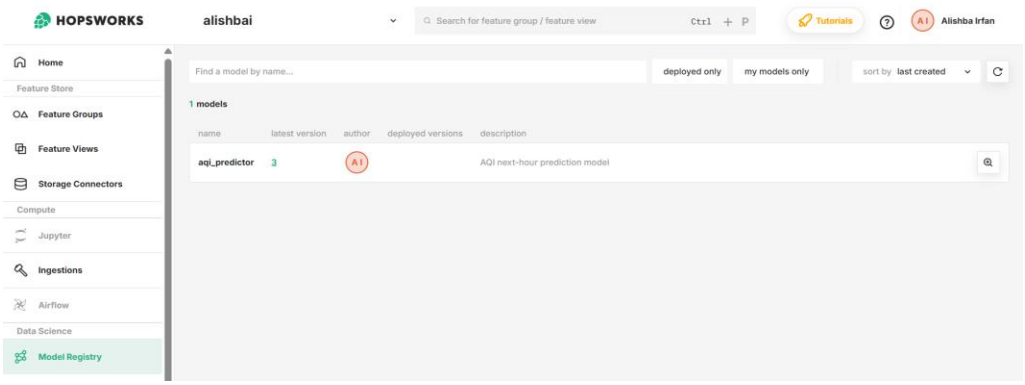
Prediction Script

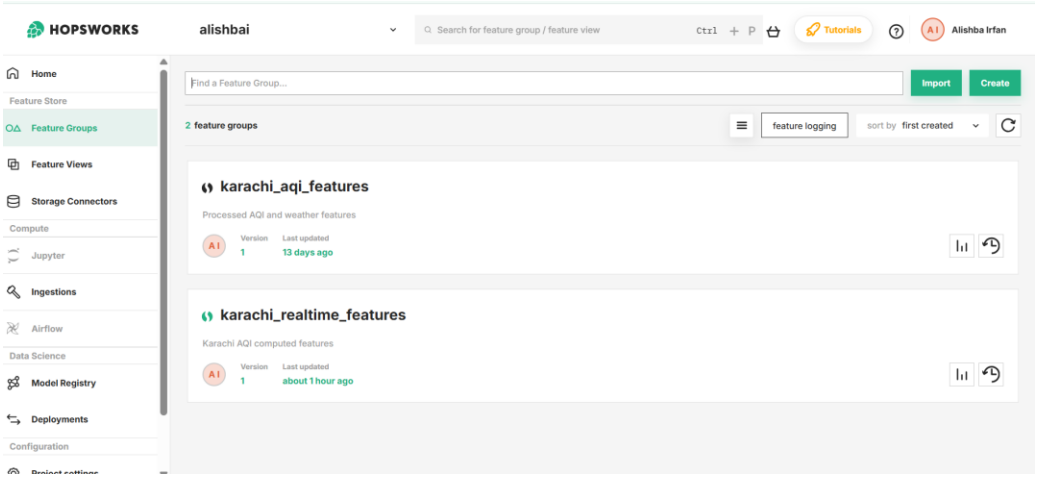
- Uses last 24 hours of real-time trained data to predict next 72 hours aqi.
- Simulated variations applied to features (temperature, pollutants) to predict future values.
- Outputs:
  - Hourly aqipredictions
  - Next 3-day average aqi

Dashboard (Streamlit)

- Next-hour AQI with alert (color-coded)
- Next 3-days AQI averages with alerts
- 72-hour AQI trend chart




HOPSWORKS
alishbai
Search for feature group / feature view
Ctrl + P
Tutorials
Alishba Irfan
Home
Feature Store
Feature Groups
Feature Views
Storage Connectors
Compute
Jupyter
Ingestions
Airflow
Data Science
Model Registry
Find a model by name...
deployed only
my models only
sort by last created
1 models
name
latest version
author
deployed versions
description
aqi_predictor
3
AQI next-hour prediction model


HOPSWORKS
alishbai
Search for feature group / feature view
Ctrl + P
Tutorials
Alishba Irfan
Home
Feature Store
Feature Groups
Feature Views
Storage Connectors
Compute
Jupyter
Ingestions
Airflow
Data Science
Model Registry
Deployments
Configuration
Find a Feature Group...
Import
Create
2 feature groups
feature logging
sort by first created
karachi_aqi_features
Processed AQI and weather features
Version
1
Last updated
13 days ago
karachi_realtime_features
Karachi AQI computed features
Version
1
Last updated
about 1 hour ago