AQI Prediction Rawalpindi

Let’s predict the Air Quality Index (AQI) in your city in the next 3 days, using a 100% server less stack.

Project Overview

The project involves the prediction of the Air Quality Index (AQI) using historical data.

Predictions were tested on **two different feature sets**:

1**. Feature Set 1** → Weather-related features + Pollutants

2**. Feature Set 2** → Only air pollutants

Models Tried:

- **Statistical Models**: ARIMA, Prophet

- **Classical ML Models**: Linear Regression, Random Forest, XGBoost, LightGBM

- **Deep Learning Models**: LSTM, RNN

Workflow:

- Initial experiments were conducted using **CSV files**.

- Later, the project was migrated to a **Hopsworks Feature Store**, which contained all historical data.

- The feature store had **two versions**:

  - **Version 1**: Feature Set 1

  - **Version 2**: Feature Set 2

Best Model:

- **LightGBM** was identified as the best-performing model for AQI prediction.

Deployment & CI/CD Pipelines:

The web application was integrated with CI/CD pipelines, consisting of **three main components**:

1. Feature Store Pipeline:

  - Runs hourly to store the latest features.

2. Training Pipeline:

  - Runs daily to retrain the model on newly added features.

3. Inference Pipeline:

- Executes immediately after the training pipeline.

- Uses the updated model to generate predictions on the latest data.

2. Data

Sources:

- Open Meto Api – [https://open-meteo.com/]

Features used:

- Pollutants: PM2.5, PM10, NO2, SO2, CO, O3

- Weather: temperature, humidity, wind speed

- Time features: hour

- Lag features: 1h, 2h, 24h back

Preprocessing:

- Scaling: StandardScaler / MinMaxScaler

- Timezone alignment (UTC → PKT)

3. EDA: (Version 1)

Columns: **['hour', 'day', 'month', 'temperature_2m (°C)',**

'relative_humidity_2m (%)', 'rain (mm)', 'wind_speed_10m (km/h)',
'wind_direction_10m (°)', 'pm10 (µg/m³)', 'pm2_5 (µg/m³)',
'carbon_monoxide (µg/m³)', 'carbon_dioxide (ppm)',
'nitrogen_dioxide (µg/m³)', 'sulphur_dioxide (µg/m³)', 'ozone (µg/m³)',
'us_aqi (USAQI)', 'hour_sin', 'hour_cos', 'is_rain']

Removed Columns: Carbon dioxide

Skewness:

rainmm 10.227251

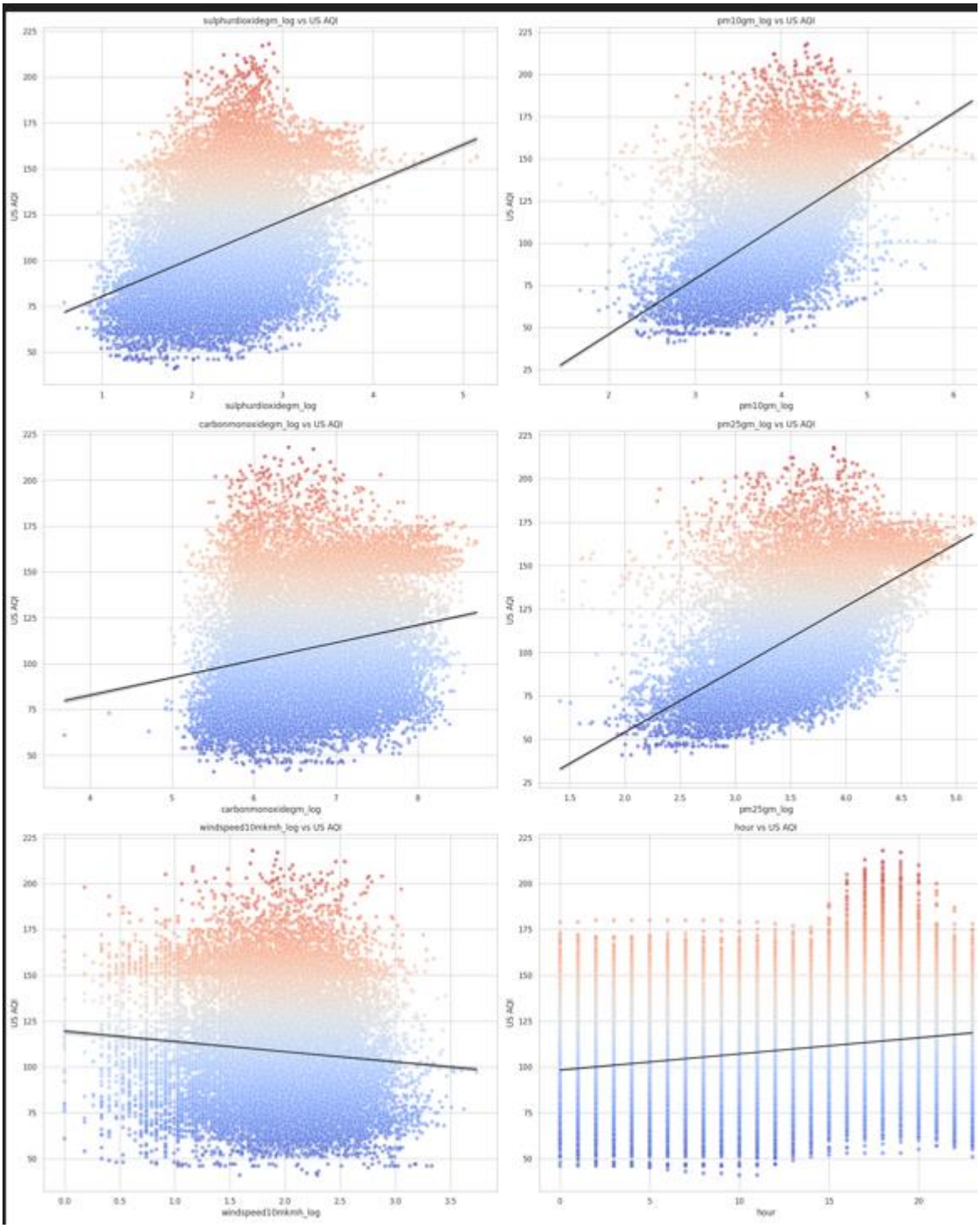
sulphurdioxidegm 3.190880

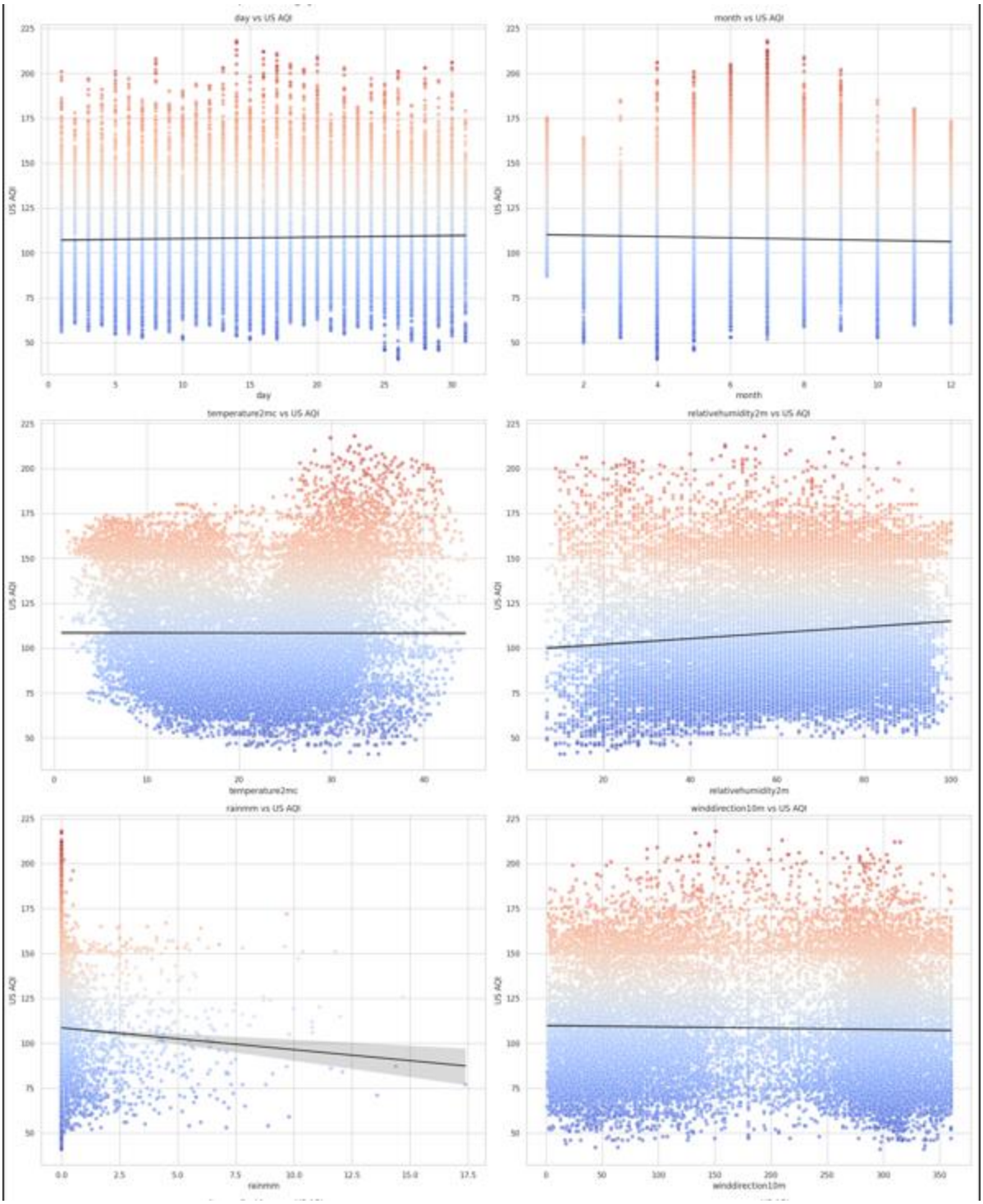
israin 2.775554

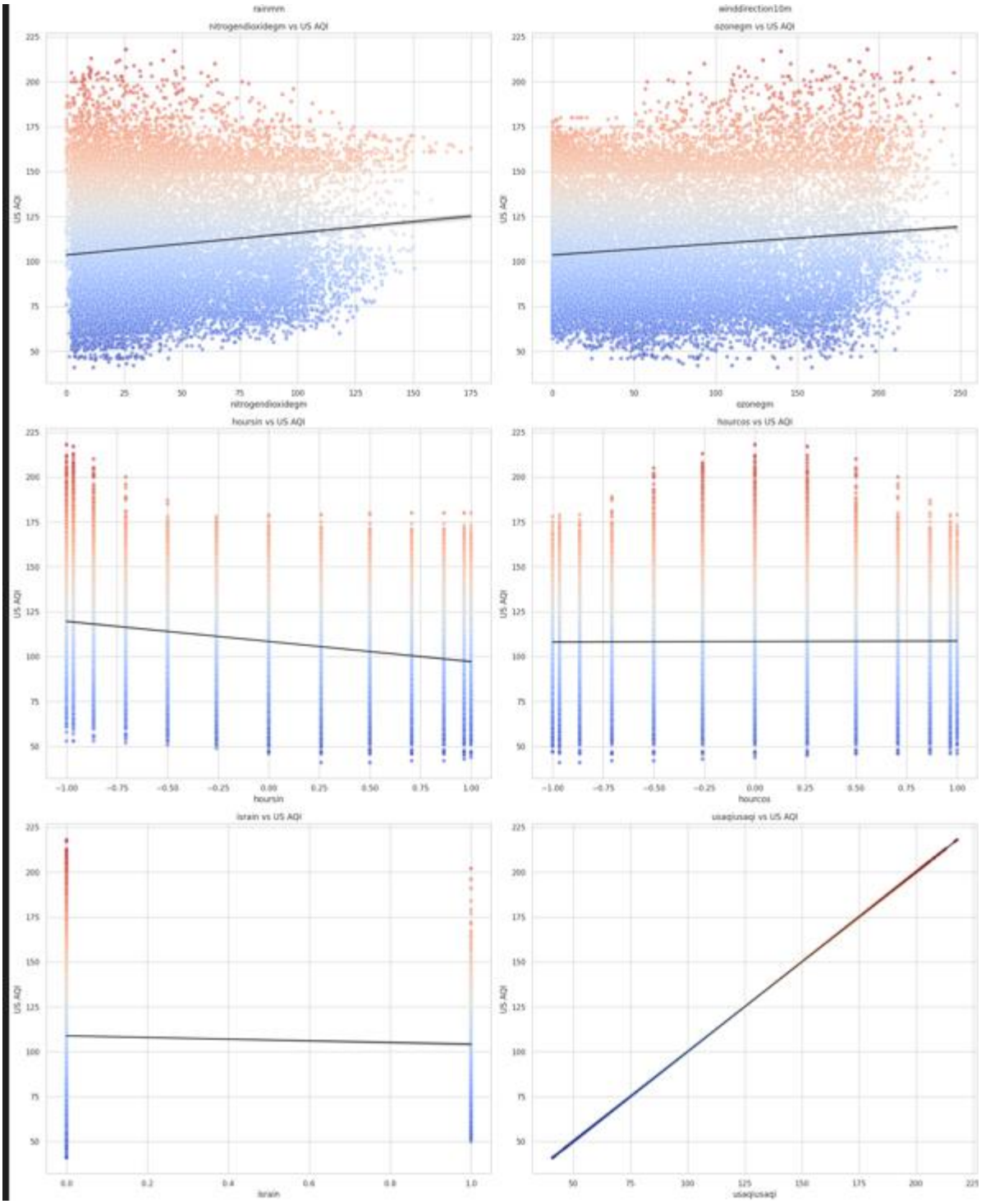
pm10gm 2.042509

carbonmonoxidegm	1.620115
pm25gm	1.476737
windspeed10mkmh	1.176434
nitrogendioxidegm	0.972093
carbondioxideppm	0.956105
ozonegm	0.526027
month	0.239558
day	0.028246
hour	0.000395
hoursin	-0.000368
hourcos	-0.000432
temperature2mc	-0.025298
winddirection10m	-0.201715
relativehumidity2m	-0.267638

Calculating Linear Relation of features with AQI:




day vs US AQI
US AQI
day
month vs US AQI
US AQI
month
temperature2mc vs US AQI
US AQI
temperature2mc
relativehumidity2m vs US AQI
US AQI
relativehumidity2m
rainmm vs US AQI
US AQI
rainmm
winddirection10m vs US AQI
US AQI
winddirection10m



Result of linear relation was that:

All the features which were related to the concentration of pollutants were having a linear relation with AQI.

Co – Relation with AQI Features:

Results:

As per the analysis I have predicted that AQI is more dependent on the pollutants values, the model predictions were not good for real time analysis. As per the results from the EDA version 2 was selected that was solely including the pollutants concentration.

Then I worked on the version 2 which was having the following features:

```
    "pm_10":
    "pm_25"
    "carbon_monoxidegm"
    "nitrogen_dioxide"
    "sulphur_dioxide"
    "Ozone"
```

4. Explaining Modals Prediction using Shap values:

Feature 0: month

Feature 1: day

Feature 2: pm10_log

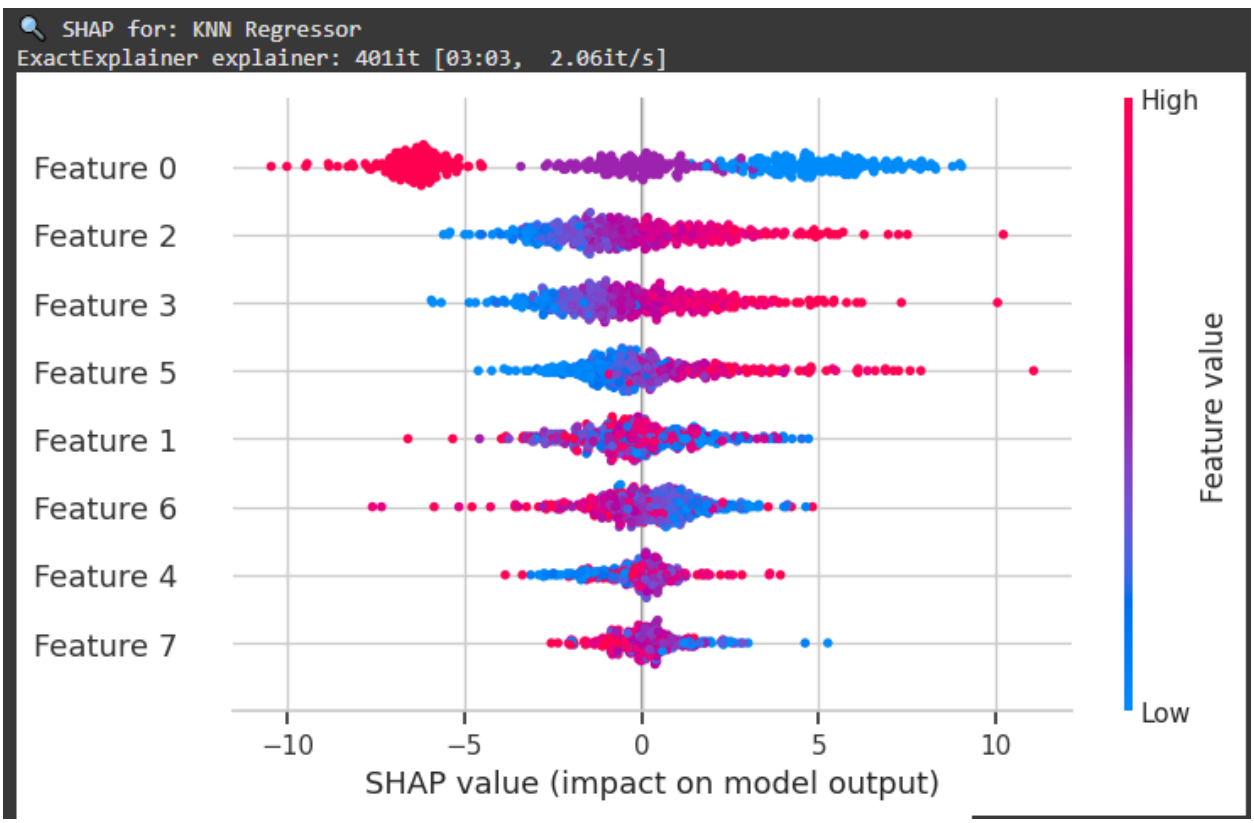
Feature 3: pm2_5_log

Feature 4: carbon_monoxide_log

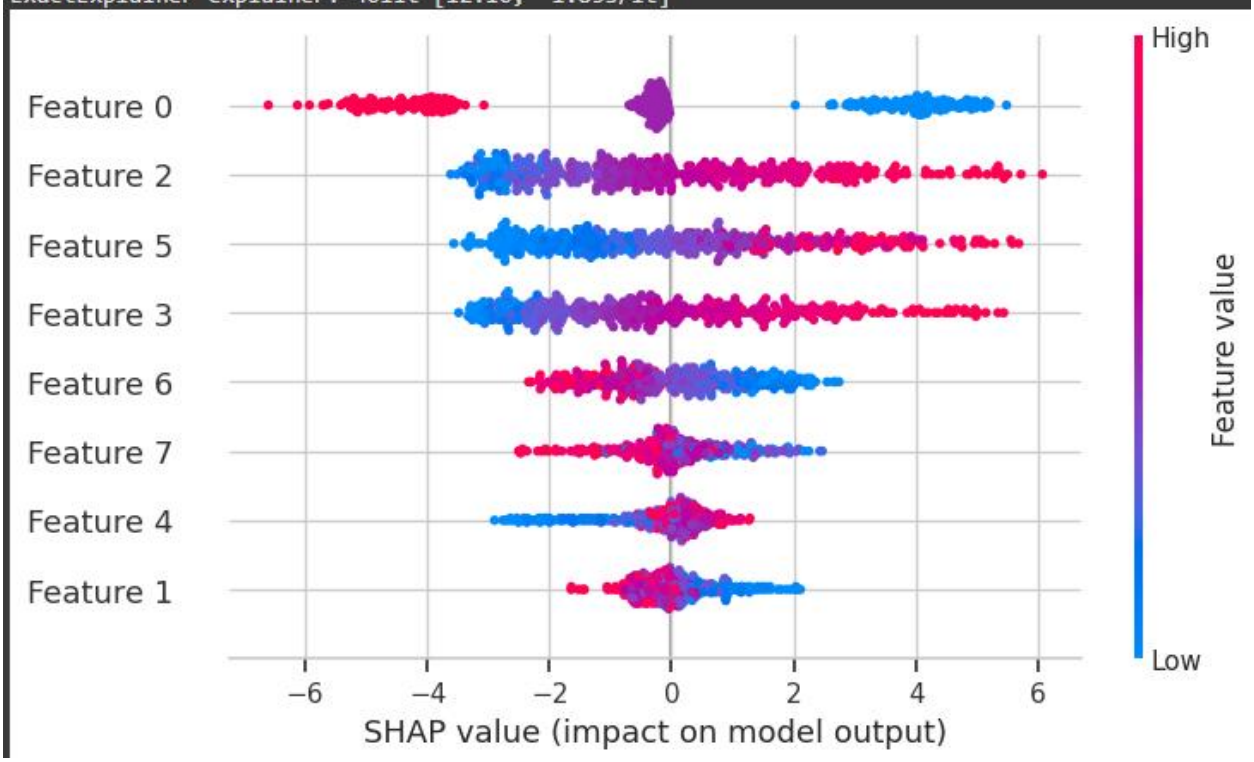
Feature 5: ozone

Feature 6: sulphur_dioxide_log

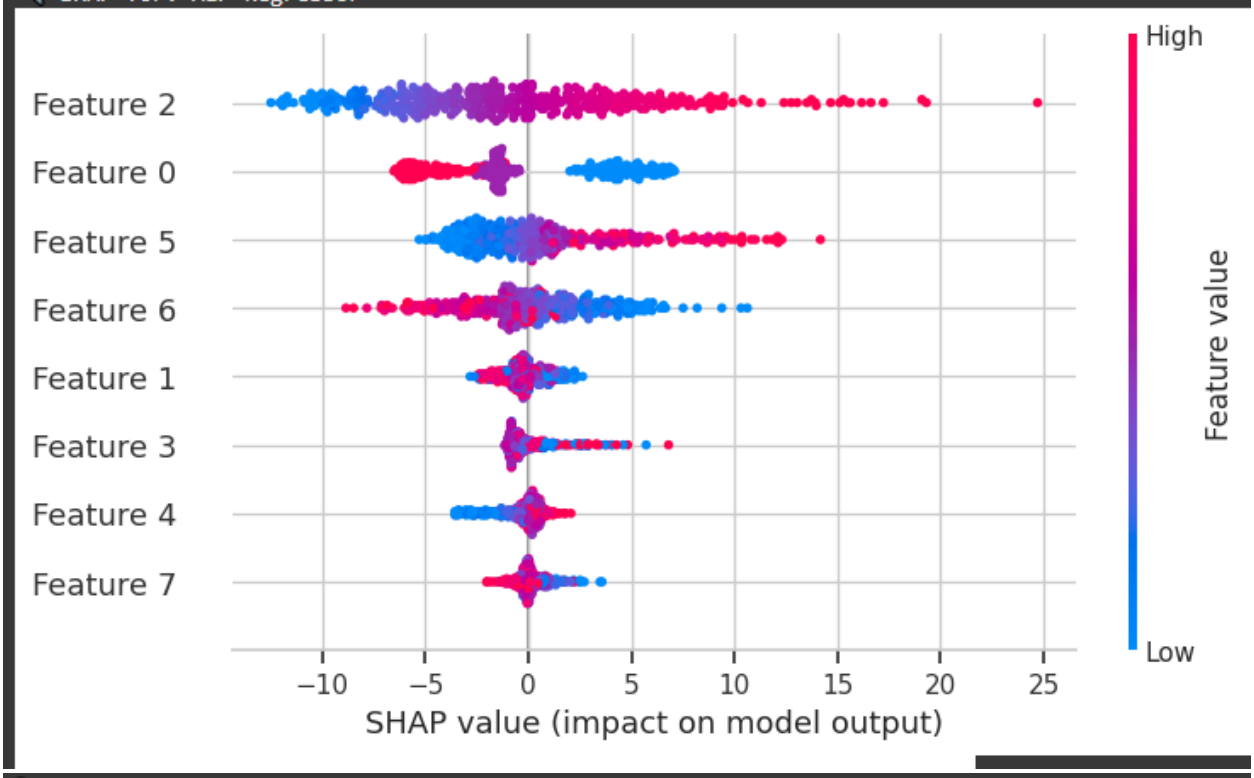
Feature 7: wind_speed_10m_log



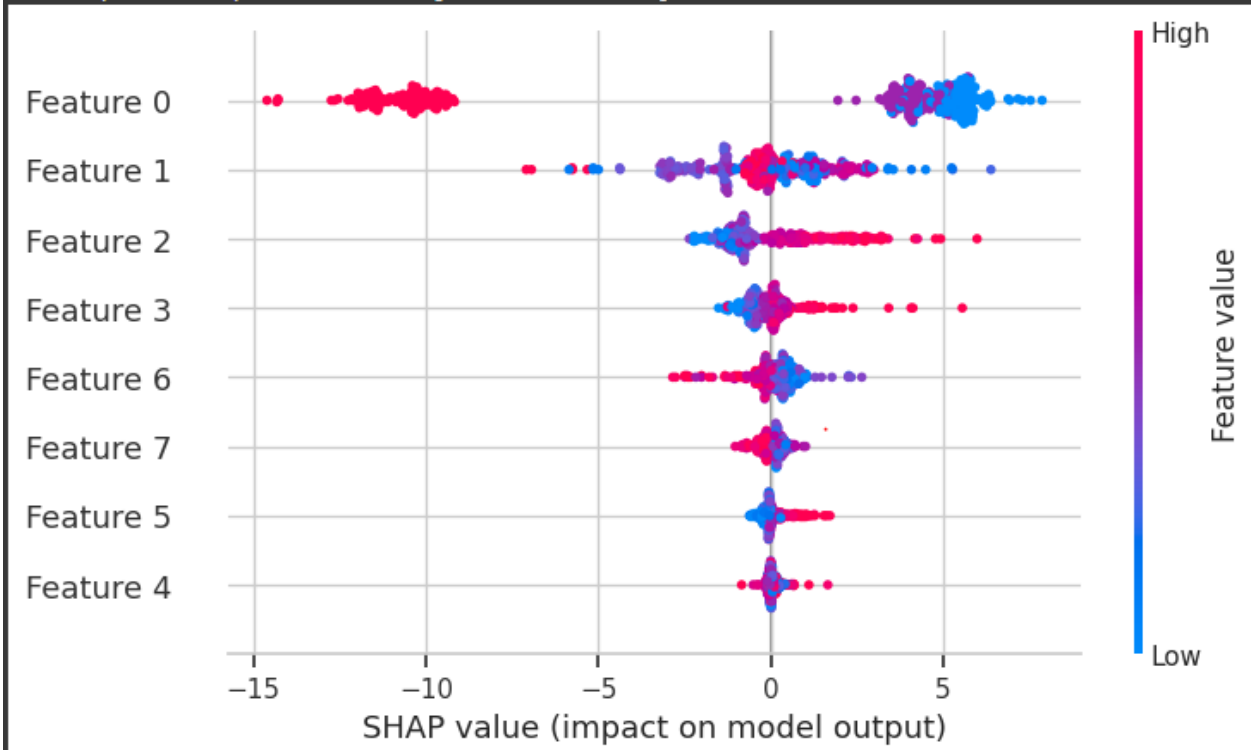
```
🔍 SHAP for: SVR
ExactExplainer explainer: 401it [12:10,  1.85s/it]
```



SHAP for: MLP Regressor



SHAP for: AdaBoost
ExactExplainer explainer: 401it [00:36, 7.90it/s]



5.Modals Metrics:

Version 2:

Statistical Models:

ARIMA:

MAE: 30.09,
RMSE: 1544.45,
R^2: -0.2713

```
=== ARIMA Forecast ===
                datetime        mean
0  2025-08-15 00:00:00   97.358120
1  2025-08-15 01:00:00  100.567657
2  2025-08-15 02:00:00  103.138915
3  2025-08-15 03:00:00  105.055568
4  2025-08-15 04:00:00  106.229798
..                 ...         ...
67 2025-08-17 19:00:00  105.713944
68 2025-08-17 20:00:00  105.697435
69 2025-08-17 21:00:00  105.713804
70 2025-08-17 22:00:00  105.697574
71 2025-08-17 23:00:00  105.713666

[72 rows x 2 columns]
```

PROPHET:

MAE: 25.22
RMSE: 955.85
R^2: 0.2132

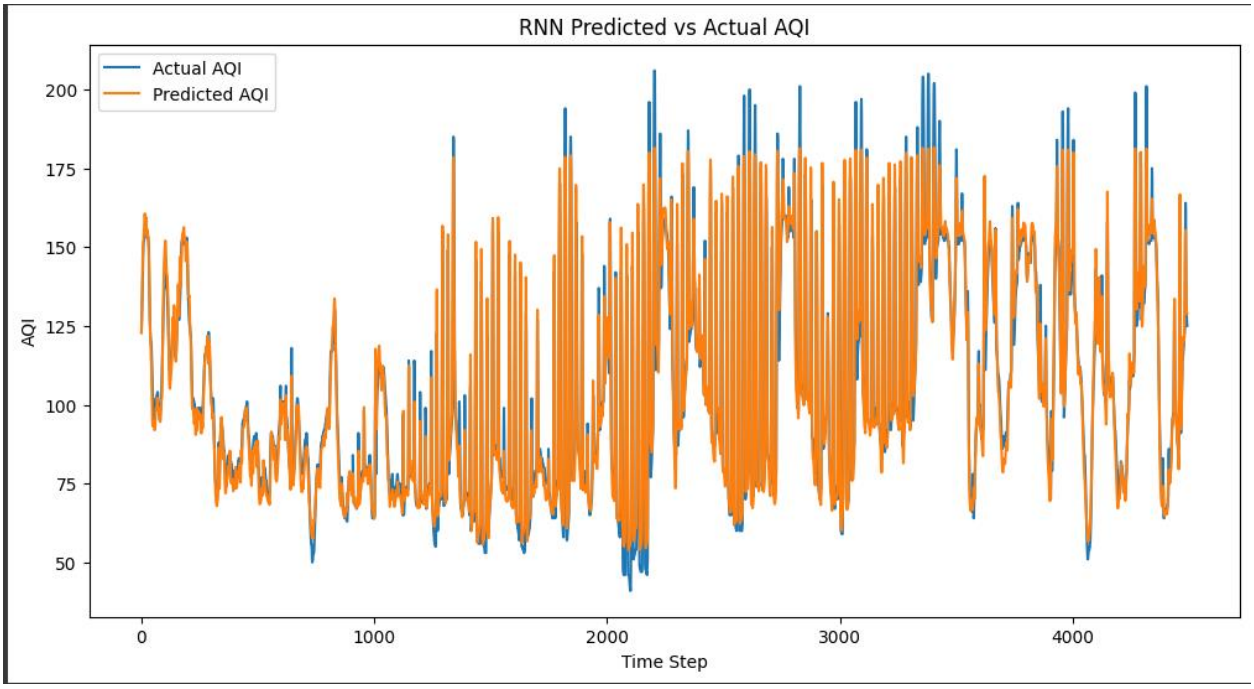
```
=== Prophet Forecast ===
                    ds   aqi  temperature  humidity  wind_speed         yhat  \
0  2025-08-15 00:00:00   6.8         24.9        90         9.9 -3344.088785
1  2025-08-15 01:00:00  10.8         24.7        90         9.8 -3398.458789
2  2025-08-15 02:00:00  15.8         24.7        91         9.5 -3436.712749
3  2025-08-15 03:00:00  16.5         24.7        92        10.1 -3446.907888
4  2025-08-15 04:00:00  14.2         24.9        92        10.5 -3415.854362
..                 ...   ...          ...       ...         ...          ...
67 2025-08-17 19:00:00  50.6         26.0        92         4.2 -2357.327089
68 2025-08-17 20:00:00  42.1         25.8        93         4.2 -2246.291826
69 2025-08-17 21:00:00  39.1         25.7        94         4.6 -2127.770840
70 2025-08-17 22:00:00  36.6         25.6        94         3.6 -2007.331069
71 2025-08-17 23:00:00  24.3         25.4        94         3.5 -1887.078195

     yhat_lower   yhat_upper
0  -3378.341285 -3309.054854
1  -3430.935798 -3364.131089
2  -3468.776189 -3404.472896
3  -3478.595703 -3412.782351
4  -3449.752123 -3382.295158
..          ...          ...
67 -2388.512723 -2325.538003
68 -2280.076311 -2212.207156
69 -2159.767653 -2095.651728
70 -2036.805993 -1973.572849
71 -1922.526424 -1854.813572

[72 rows x 8 columns]
```

Deep Learning Models:

RNN:

RMSE: 5.57

R^2 Score: 0.97

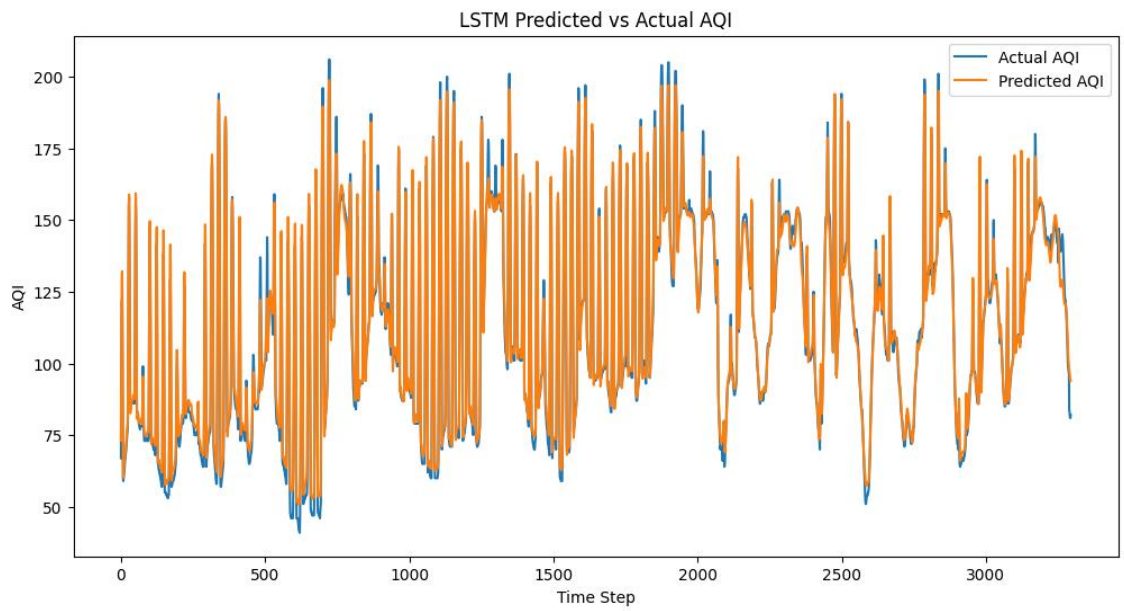


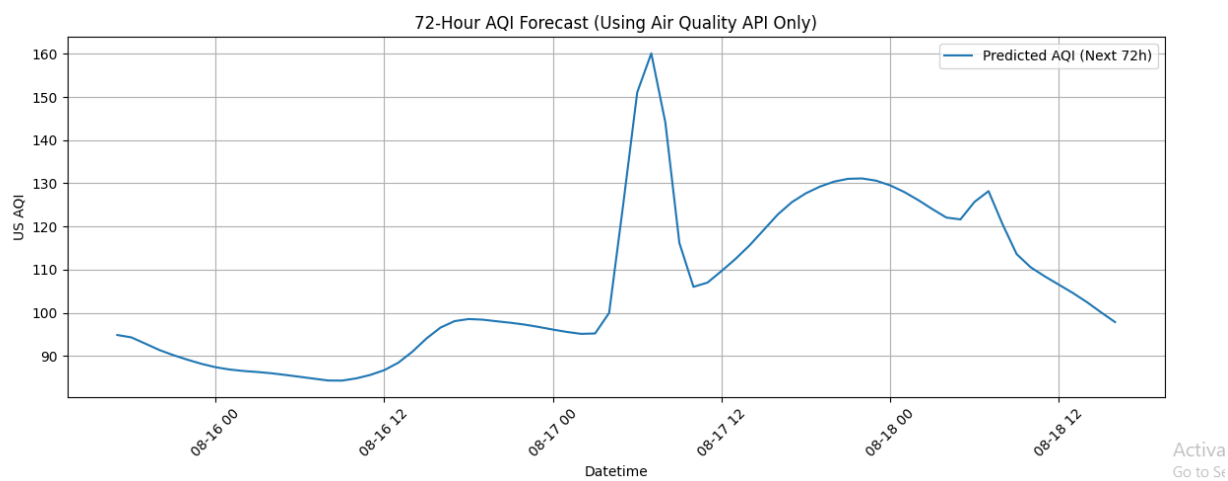
	datetime	predicted_us_aqi
0	2025-08-04 02:00:00+00:00	127.970711
1	2025-08-04 03:00:00+00:00	128.254410
2	2025-08-04 04:00:00+00:00	129.883347
3	2025-08-04 05:00:00+00:00	131.752136
4	2025-08-04 06:00:00+00:00	133.163925

Were very away from the real values.

LSTM:

RMSE: 4.07
R^2 Score: 0.99





Machine Learning Modals:

Random Forest:

MAE: 6.44
RMSE: 77.77
R^2: 0.9362

Correlation with the Open Meto API values:

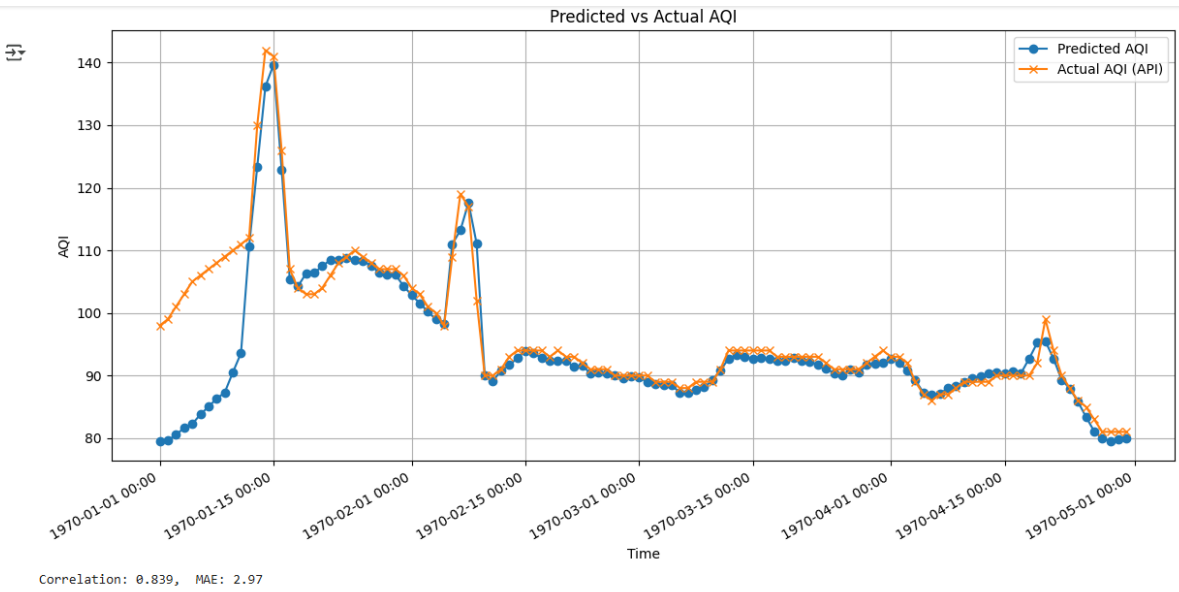
```
[info] Plot saved to: rf_aqi_artifacts/plots/xgb_72h_pred_vs_actual.png

[summary] 72h window alignment metrics (API us_aqi vs XGB prediction):
  MAE:  5.98
  RMSE: 55.81
  Corr: 0.9066

✅ Done. Artifacts are in: /content/rf_aqi_artifacts
```

XG BOOST:

MAE: 2.56

RMSE: 17.27

R^2: 0.9858

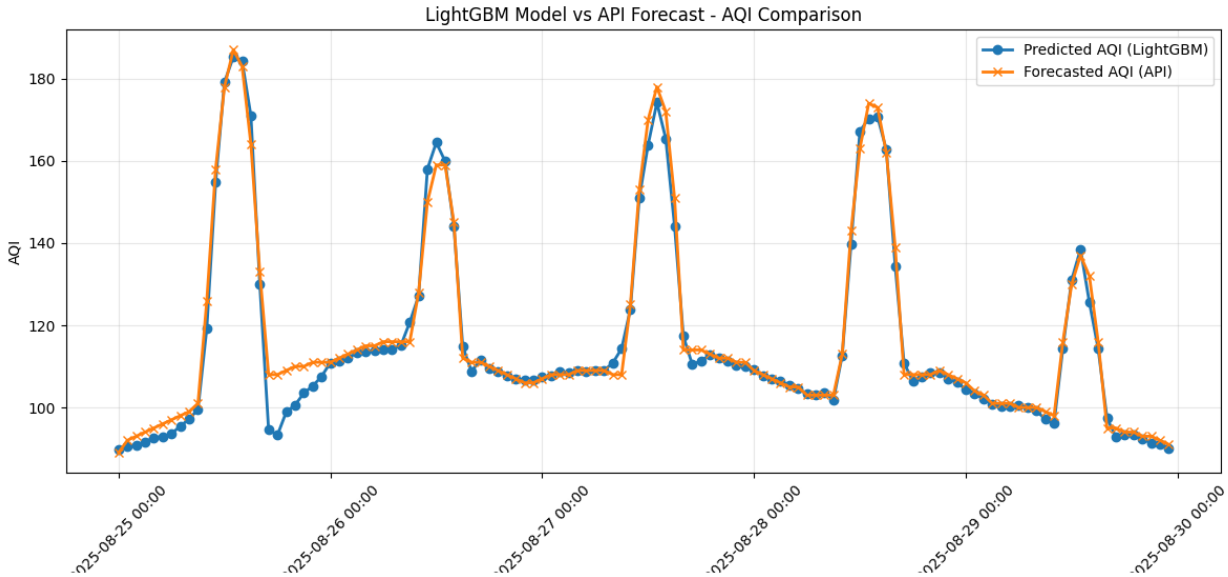


Light GBM: (The best Model)

MAE: 2.50,

RMSE: 25.14,

R^2: 0.9791



6. Final Application:

Share

Dashboard Controls

Data Source

Feature Store

Auto-refresh (5 min)

Advanced Options

Records loaded: 159

Data source: Feature Store

Debug: Show column names

Rawalpindi AQI Forecast Dashboard

Real-time **74-hour AQI forecasts** powered by LightGBM ML model

Location: Rawalpindi, Punjab, Pakistan (33.5973°N, 73.0479°E)

Total Forecasts	Average AQI	AQI Range	Model Version
159	100	63 - 202	Unknown

Current Forecasted AQI

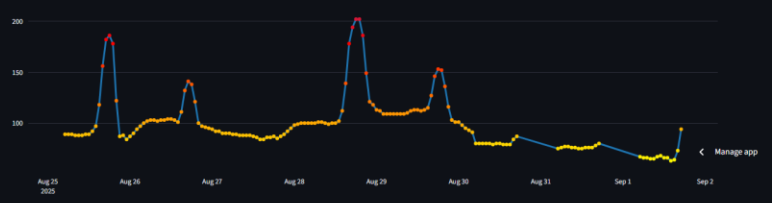
100

Moderate

Forecasted for 2025-08-28 12:00:00 PKT

Manage app

74-Hour AQI Forecast



Manage app

Detailed Forecast Table

Forecast Time (PKT)	AQI Value	AQI Category
2025-08-25 05:00:00 PKT	89	Moderate
2025-08-25 06:00:00 PKT	89	Moderate
2025-08-25 07:00:00 PKT	89	Moderate
2025-08-25 08:00:00 PKT	88	Moderate
2025-08-25 09:00:00 PKT	88	Moderate
2025-08-25 10:00:00 PKT	88	Moderate
2025-08-25 11:00:00 PKT	89	Moderate
2025-08-25 12:00:00 PKT	89	Moderate
2025-08-25 13:00:00 PKT	92	Moderate
2025-08-25 14:00:00 PKT	97	Moderate

Download Forecast Data (CSV)

Powered by LightGBM ML Model | Data from Hopsworks Feature Store

Location: Rawalpindi, Punjab, Pakistan (33.5973°N, 73.0479°E)

Last updated: 2025-08-28 12:18:12 PKT

Showing 159 forecast points | All times in Pakistan Time (PKT)

Manage app