AIR QUALITY INDEX IN KARACHI FOR THE NEXT 3 DAYS

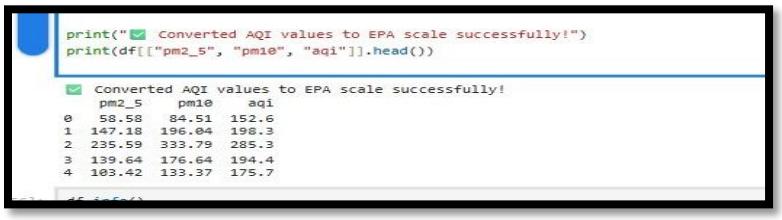
Data Collection:

We collected historical AQI and weather data from OpenWeather API over the past year. Unlike typical hourly or daily data, we chose a **5-hour interval**: Early in the data exploration, consecutive AQI values were often identical. Using such repeated values could mislead the model into memorizing these values (overfitting), reducing its predictive ability During data collection from the OpenWeather API, the following features were retrieved for each timestamp:

Pre-Processing:

Converted to US EPA AQI Scale:

The AQI values we fetched from OpenWeather were on a simple 0–5 scale, which is too basic for real-world use and doesn't match international standards. To make the data meaningful and comparable, we converted it to the US EPA AQI scale, which ranges from 0 to 500.



```
print("✅ Converted AQI values to EPA scale successfully!")
print(df[["pm2_5", "pm10", "aqi"]].head())

✅ Converted AQI values to EPA scale successfully!
    pm2_5    pm10    aqi
0   58.58   84.51  152.6
1  147.18  196.04  198.3
2  235.59  333.79  285.3
3  139.64  176.64  194.4
4  103.42  133.37  175.7
```

Here's the approach we used:

1. Separate calculations for PM2.5 and PM10:
2. We first looked at the concentration of PM2.5 (fine particles) and PM10 (coarse particles).
3. Each concentration was mapped to its corresponding range on the EPA scale. This tells us how harmful the air quality is based on the pollutant amount.
4. Choosing the worst-case AQI
5. After converting both PM2.5 and PM10 to the EPA scale, we took the higher of the two as the final AQI.
6. This ensures that the AQI reflects the worst air quality, which is important for safety and public health.

Result: The converted AQI now ranges from 0 to 500, following the EPA standard. This scaled AQI can now be used reliably for analysis, modeling, and forecasting.

Handling Missing and Duplicate Values:

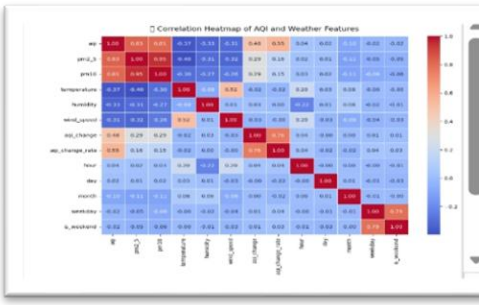
1. Missing timestamps or duplicate rows were removed.
2. Timestamps were converted to datetime objects and sorted chronologically.

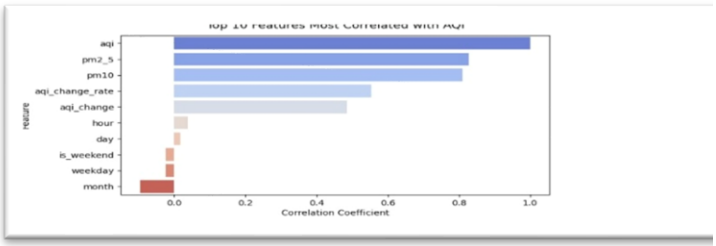
Feature Engineering:

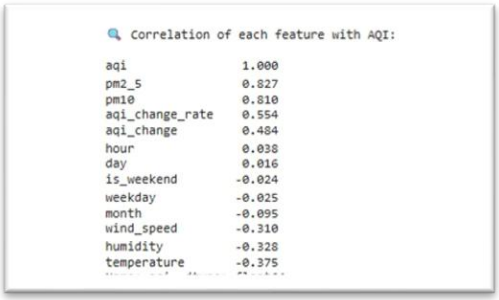
To enhance model performance, we derived both time-based and AQI-based features:
Time-based features: Hour, day, month, weekday, and whether it is a weekend.
AQI-derived features:

- aqi_change: Absolute change in AQI from previous measurement.
- aqi_change_rate: Relative change, capturing volatility in air quality

Correlation between AQI and features:





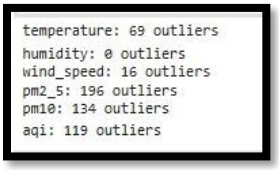


Correlation of each feature with AQI:

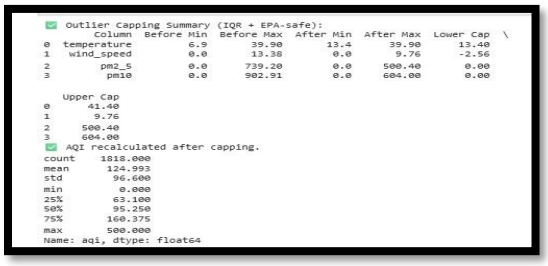
aqi	1.000
pm2_5	0.827
pm10	0.810
aqi_change_rate	0.554
aqi_change	0.484
hour	0.038
day	0.016
is_weekend	-0.024
weekday	-0.025
month	-0.095
wind_speed	-0.310
humidity	-0.328
temperature	-0.375

OUTLIER DETECTION:

We handled outliers by checking which values were much lower or higher than most of the data, using a standard statistical method. Any extreme values that fell outside this typical range were capped to keep the dataset consistent and safe for modeling.



temperature: 69 outliers
humidity: 0 outliers
wind_speed: 16 outliers
pm2_5: 196 outliers
pm10: 134 outliers
aqi: 119 outliers



Outlier Capping Summary (IQR + EPA-safe):

	Column	Before Min	Before Max	After Min	After Max	Lower Cap	Upper Cap
0	temperature	6.9	39.90	13.4	39.90	13.40	41.40
1	wind_speed	0.0	13.38	0.0	9.76	-2.56	9.76
2	pm2_5	0.0	739.20	0.0	500.40	0.00	500.40
3	pm10	0.0	902.91	0.0	604.00	0.00	604.00

AQI recalculated after capping.

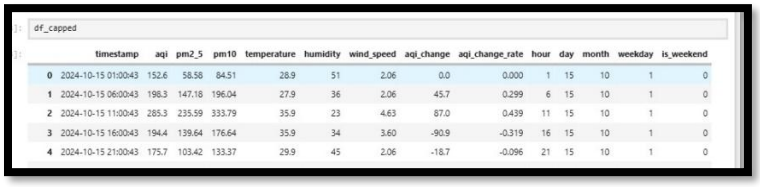
count	1818.000
mean	124.993
std	96.600
min	0.000
25%	63.100
50%	95.250
75%	160.375
max	500.000

Name: aqi, dtype: float64

CAPPING Technique for Outliers:

We handled outliers by checking which values were much lower or higher than most of the data, using a standard statistical method. Any extreme values that fell outside this typical range were capped to keep the dataset consistent and safe for modeling.

Now we have processed dataframe df_capped that we will use for training

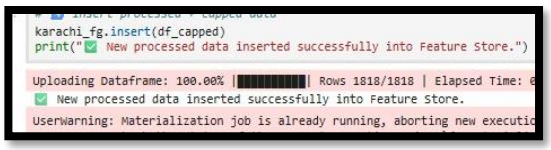


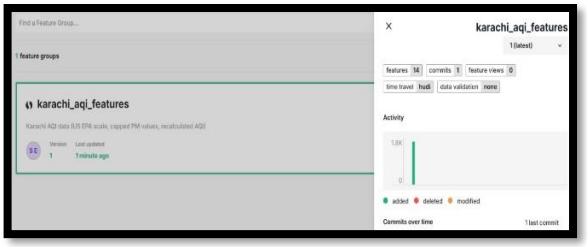
df_capped

	timestamp	aqi	pm2_5	pm10	temperature	humidity	wind_speed	aqi_change	aqi_change_rate	hour	day	month	weekday	is_weekend
0	2024-10-15 01:00:43	152.6	58.58	84.51	28.9	51	2.06	0.0	0.000	1	15	10	1	0
1	2024-10-15 06:00:43	198.3	147.18	196.04	27.9	36	2.06	45.7	0.299	6	15	10	1	0
2	2024-10-15 11:00:43	285.3	235.59	333.79	35.9	23	4.63	87.0	0.439	11	15	10	1	0
3	2024-10-15 16:00:43	194.4	139.64	176.64	35.9	34	3.60	-90.9	-0.319	16	15	10	1	0
4	2024-10-15 21:00:43	175.7	103.42	133.37	29.9	45	2.06	-18.7	-0.096	21	15	10	1	0

Feature Storage:

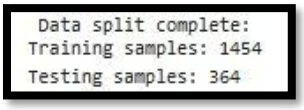
Cleaned and processed features are stored in Hopsworks Feature Store for reproducibility and easy access.





Model Training:

The dataset is split into 80% training and 20% testing for model evaluation.



We experimented with both parametric and non-parametric models, as well as a neural network:

- *Non-parametric*: **Random Forest**, Extra Trees, KNN, XGBoost, SVR
- *Parametric:* Linear Regression, Ridge, Lasso, Elastic Net
- *Neural Network:* ANN

How we selected the best model out of all the algorithm's best models? To ensure models perform well and generalize, we defined a combined score:

$$\textbf{Combined Score} = \text{Test RMSE} + \lambda \times |\text{Train RMSE} - \text{Test RMSE}|$$

. λ weighs overfitting risk. The metric rewards low test RMSE while penalizing models that overfit training data.

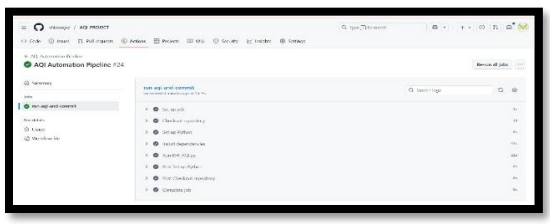
Random Forest as Best Model:

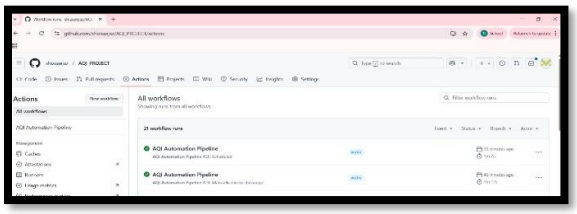
Multiple hyperparameter configurations were tested. Random Forest achieved the lowest combined score and was selected as the final model. The model is saved as **best_random_forest.pkl.***

Prediction and Deployment:

Using the best model, AQI predictions are generated:
forecast_data = get_forecast(df_capped, model_path="best_random_forest.pkl", save_csv=True)
Predictions are saved to **latest_predictions.csv** and integrated into **a Streamlit frontend**.

CI/CD via **GitHub Actions** automatically fetches new data and **updates predictions every 5 hours**.





SkyCast

Air Quality Index Prediction for Karachi

Karachi, Pakistan

03:10 PM

437

Hazardous

HEALTH RECOMMENDATIONS

- Avoid outdoor activities
- Stay indoors with air purifier
- Wear N95 if absolutely necessary

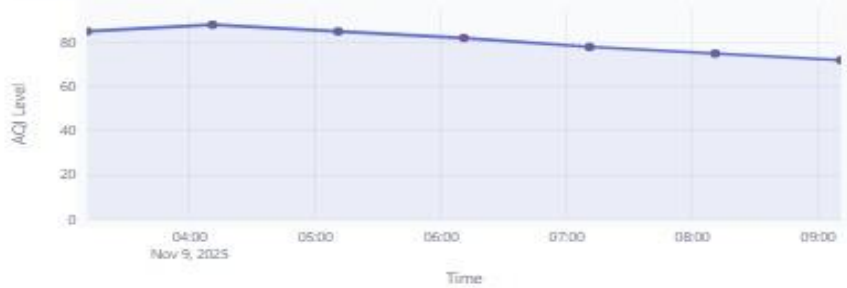
Current Weather

TEMPERATURE	HUMIDITY	WIND SPEED
23°C	44%	2 km/h
Partly Cloudy	Moderate	Light Breeze

3-Day AQI Forecast

Today	Tomorrow	Day After
Nov 09, 2025	Nov 10, 2025	Nov 11, 2025
MIN 432 / MAX 442	MIN 432 / MAX 442	MIN 432 / MAX 442

AQI Trend Over Time



SkyCast - Real-time Air Quality Predictions

Last updated: 2025-11-09 15:11:00