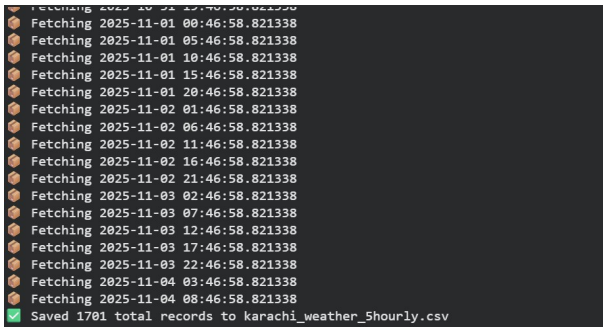
AIR QUALITY INDEX IN KARACHI FOR THE NEXT 3 DAYS

INTRODUCTION:

This project aims to predict the Air Quality Index (AQI) using machine learning techniques. The system analyzes environmental pollutant data such as PM2.5, PM10, NO_2, and CO levels to forecast air quality in Karachi. The goal is to provide an efficient and data-driven approach to monitor air pollution and raise awareness about air quality conditions.

Data Collection and Sampling Strategy:

For this project, historical AQI and weather data were collected using the OpenWeather API at **5-hour intervals** over the course of **one year**. Collecting data every 5 hours, instead of hourly, ensures meaningful variation between data points and helps **reduce the risk of overfitting** by avoiding redundant, highly similar records.




```
df=pd.read_csv("karachi_weather_5hourly.csv")
df.head()
```

	timestamp	aqi	pm2_5	pm10	temperature	humidity	wind_speed
0	2024-11-04 18:46:58	4	67.15	124.15	24.9	78	0.00
1	2024-11-04 23:46:58	4	58.27	96.45	24.9	83	2.06
2	2024-11-05 04:46:58	5	265.79	327.71	28.9	65	2.06
3	2024-11-05 09:46:58	3	37.26	71.72	33.9	24	5.14
4	2024-11-05 14:46:58	3	46.13	81.12	27.9	44	2.57

Exploratory Data Analysis (EDA) :

During EDA, we first **checked for duplicates and outliers**. Since weather and air quality data are naturally variable, extreme values could represent genuine spikes rather than errors. To preserve the **originality of the dataset**, we capped the outliers instead of removing them.

```
temperature: 69 outliers
humidity: 0 outliers
wind_speed: 17 outliers
pm2_5: 177 outliers
pm10: 119 outliers
aqi: 93 outliers
```

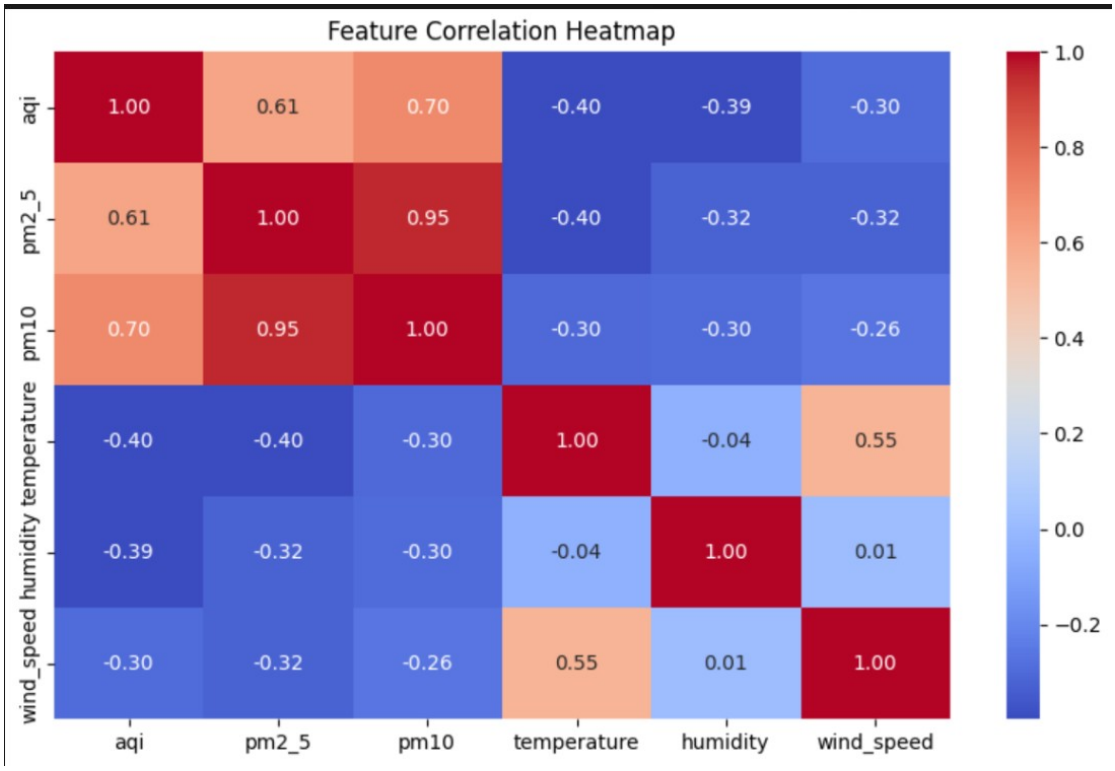
```
After capping:
       temperature     humidity   wind_speed        pm2_5         pm10  \
count  1701.000000  1701.000000  1701.000000  1701.000000  1701.000000
mean     26.957166    55.576720     3.976902    50.331931   110.406470
std       5.710172    21.768586     2.153663    47.278369    73.900155
min      13.400000     5.000000     0.000000     2.260000     6.310000
25%      23.900000    39.000000     2.060000    16.820000    58.640000
50%      27.900000    58.000000     3.600000    29.880000    88.770000
75%      30.900000    74.000000     5.140000    70.650000   148.150000
max      39.900000   100.000000     9.760000   151.395000   282.415000

              aqi
count  1701.000000
mean      3.508230
std       1.094877
min       1.500000
25%       3.000000
50%       3.000000
75%       4.000000
max       5.000000
```

```
print("\nAfter capping:")
print(df[numeric_cols])
```

```
...
    After capping:
          temperature  humidity  wind_speed  pm2_5    pm10  aqi
    0            22.9      88.0        1.54  48.50   76.82  3.0
    1            33.9      43.0        5.14  31.09   71.04  3.0
    2            29.9      62.0        3.60  29.57   73.40  3.0
    3            26.9      83.0        2.06  60.57  106.61  4.0
    4            26.9      83.0        3.09  35.62   67.14  3.0
    ...           ...       ...         ...    ...     ...  ...
    1696         14.9      38.0        2.57  50.01  128.00  4.0
    1697         30.9      10.0        1.54  57.80  132.41  4.0
    1698         30.9      19.0        4.12  67.93  135.84  4.0
    1699         23.9      57.0        2.57  44.79  101.98  4.0
    1700         16.9      88.0        2.06  36.04   91.47  3.0
```

After addressing these issues, we examined **feature relationships**, which revealed that **PM2.5 and PM10** are most closely related to AQI and have the greatest influence on it.



Feature Selection:

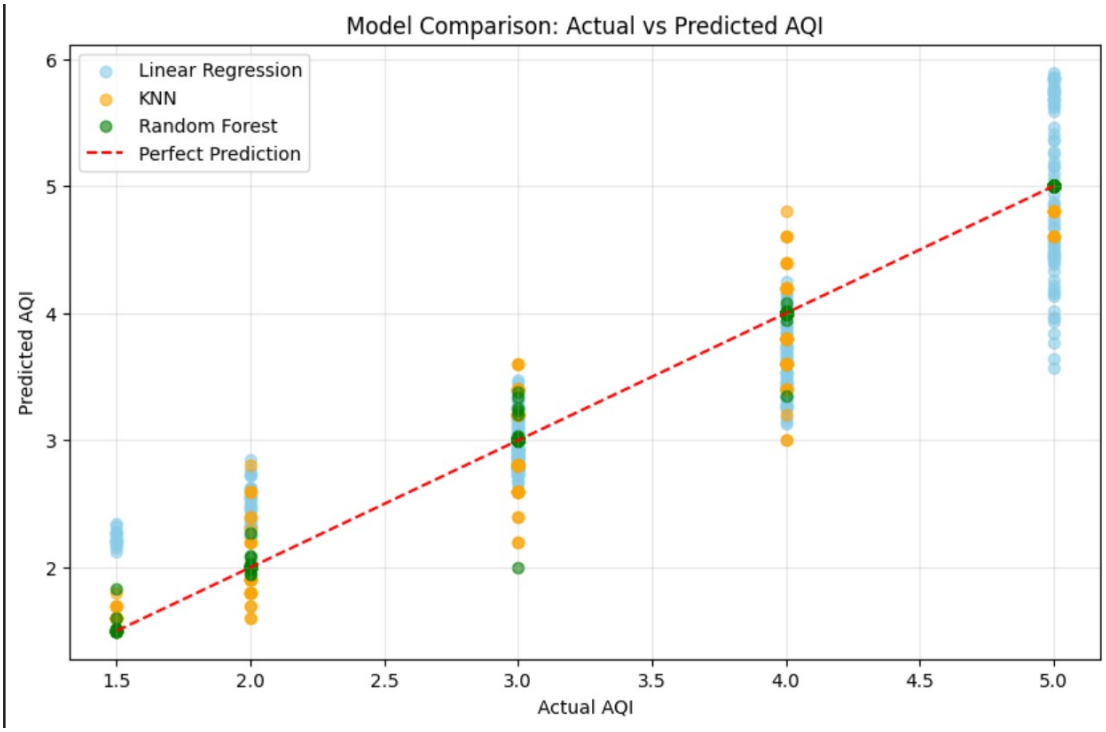
During feature selection, we analyzed the correlation between variables and AQI. **PM2.5 and PM10** showed a strong positive correlation with each other and with AQI, indicating that particulate matter is a major factor influencing air quality. **Temperature** had a moderate negative correlation with AQI, **humidity** had a weak negative correlation with pollutants, and **wind speed** was positively correlated with temperature and slightly negatively correlated with pollutants, suggesting it can help disperse pollutants. Despite these correlations, we retained all features except **timestamp**, which was dropped to simplify the dataset.

After selecting the features, we applied **MinMaxScaler** to normalize all variables to a common scale. This ensures that no single feature dominates the model due to its magnitude and helps the model **learn efficiently and converge faster**.

Model Selection and Training:

We selected **Linear Regression, KNN Regression, and Random Forest Regression** to model AQI. The data was trained on all three models, and a **sanity check** was performed to ensure that the models performed consistently on both training and test datasets. Among all three, **Random Forest Regression performed the best**, producing the most accurate and reliable predictions on both the predicted and actual data.

```
•••  Random Forest Regressor Results:
     R² Score: 0.986
     Mean Absolute Error: 0.029
     Root Mean Squared Error: 0.131
```



Deployment and Production Workflow:

The final Random Forest model was deployed on **Streamlit Cloud**, where it fetches the latest **5 days of data** via the OpenWeather API and generates AQI predictions in real time. Both the **predicted data and the trained model** are uploaded to **Hopsworks**, ensuring that the system maintains an up-to-date dataset and model for further analysis and production use.