INTRODUCTION:

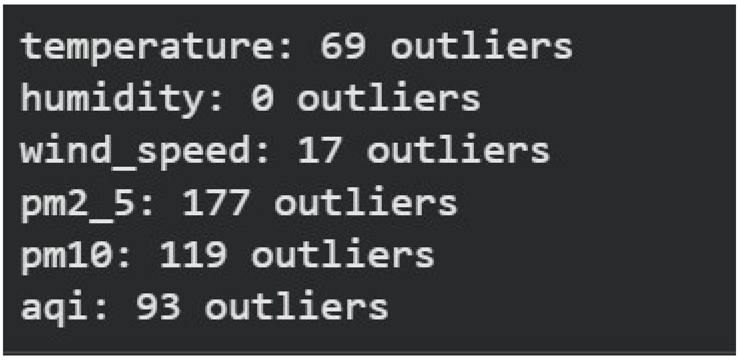
This project aims to predict the Air Quality Index (AQI) using machine learning techniques. The system analyzes environmental pollutant data such as PM2.5, PM10, NO_2, and CO levels to forecast air quality in Karachi. The goal is to provide an efficient and data-driven approach to monitor air pollution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

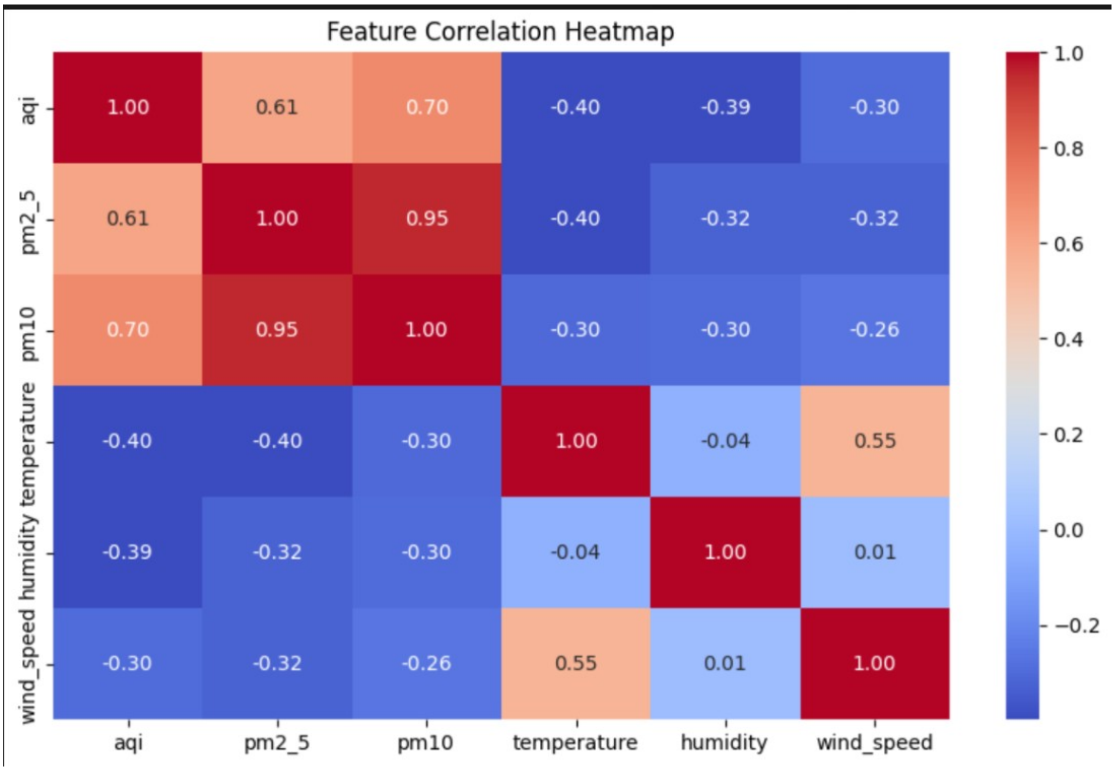
Exploratory Data Analysis (EDA) :

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 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 Correlation Heatmap

	aqi	pm2_5	pm10	temperature	humidity	wind_speed
aqi	1.00	0.61	0.70	-0.40	-0.39	-0.30
pm2_5	0.61	1.00	0.95	-0.40	-0.32	-0.32
pm10	0.70	0.95	1.00	-0.30	-0.30	-0.26
temperature	-0.40	-0.40	-0.30	1.00	-0.04	0.55
humidity	-0.39	-0.32	-0.30	-0.04	1.00	0.01
wind_speed	-0.30	-0.32	-0.26	0.55	0.01	1.00

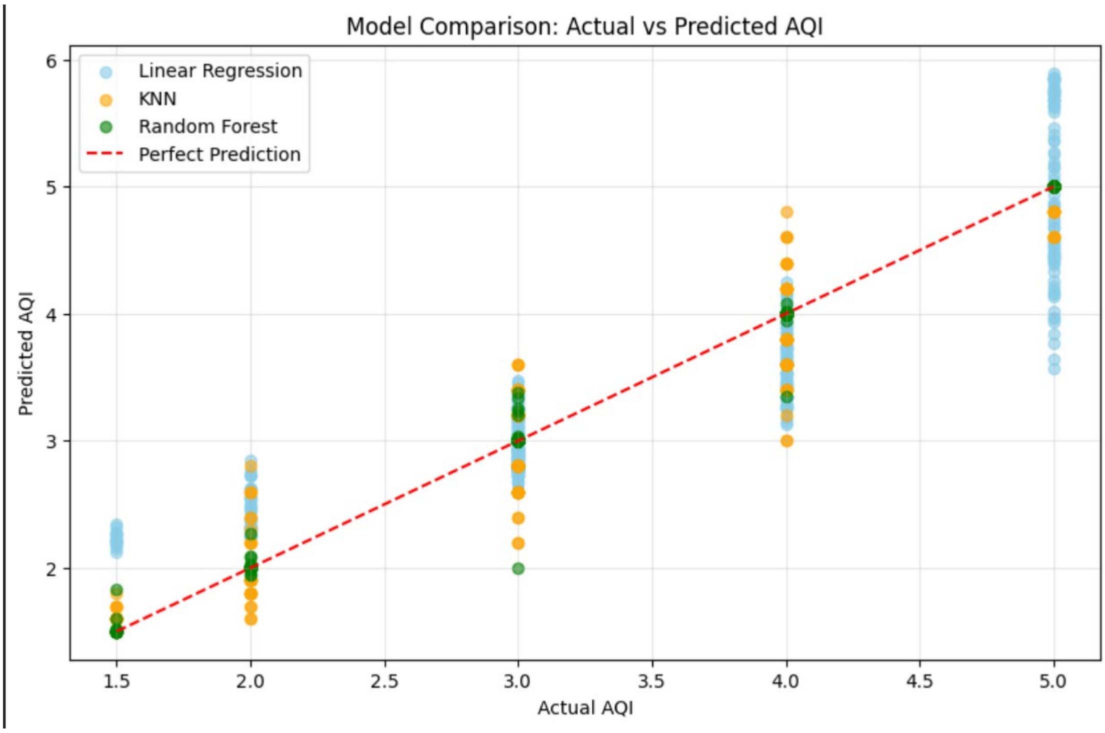
Feature Selection:

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:

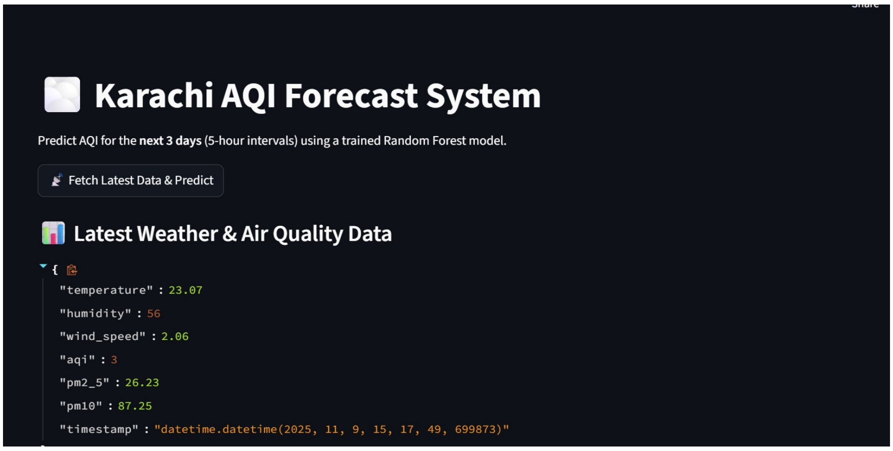
We selected **Linear Regression, KNN Regression, and Random Forest Regression** to model AQI. The data was trained on all three models, and a **sanity check** was performed to ensure that the models performed consistently on both training and test datasets. Among all three, **Random Forest Regression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



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.



Karachi AQI Forecast System

Predict AQI for the **next 3 days** (5-hour intervals) using a trained Random Forest model.

Latest Weather & Air Quality Data

```
{
  "temperature" : 23.07
  "humidity" : 56
  "wind_speed" : 2.06
  "aqi" : 3
  "pm2_5" : 26.23
  "pm10" : 87.25
  "timestamp" : "datetime.datetime(2025, 11, 9, 15, 17, 49, 699873)"
```