Phase-3 Submission

Student Name: Dhyanesh V

Register Number: 712523104018

Institution: PPG Institute Of Technology

Department:B.E Computer Science and Engineering

Date of Submission: 16/05/2025

Github Repository Link: AirPreQ

● Problem Statement

Accurate and real-time prediction of Air Quality Index (AQI) is essential for public health safety and informed policymaking. This project focuses on developing a supervised regression model to estimate AQI levels using various environmental pollutant indicators and deploying it as a user-friendly web application. It also supports analysis of user-uploaded datasets and live API-based air quality predictions.

Problem Type: Supervised Regression

● Abstract

Air pollution poses one of the most significant environmental and public health challenges in today's world, especially in rapidly urbanizing nations like India. The Air Quality Index (AQI) is a crucial measure used to assess and communicate how polluted the air currently is or how polluted it is forecast to become. However, real-time monitoring stations are limited, and manually analyzing pollutant data to estimate AQI is time-consuming and prone to error. To address this, we propose a comprehensive machine learning-based solution for predicting AQI levels using various air pollutant concentrations such as PM2.5, PM10, NO_2, CO, SO_2, O_3, and more.

This project leverages data science and artificial intelligence to predict AQI values with high accuracy using supervised regression techniques, primarily focusing on Random Forest Regressor due to its robustness and capability to handle complex relationships between multiple pollutants. Our system is trained on publicly available datasets and is designed to be scalable and easily extendable with new data sources.

The application features a user-friendly web interface built using Flask, HTML, CSS, and JavaScript, and is deployed on a cloud platform (Render) to ensure platform independence and accessibility across devices. The web app allows users to explore AQI predictions in three different modes: using a predefined dataset, by uploading their own custom dataset, or through real-time data fetched from the IQAir API. The system also provides meaningful visualizations such as trend lines, correlation heatmaps, and pollutant contribution charts to enhance interpretability.

From a software engineering perspective, the backend is modularized to handle preprocessing, model training, evaluation, and visualization in a structured manner, promoting easy maintenance and scalability. Security practices are followed by storing sensitive keys such as the API token in environment variables rather than hardcoding.

This project not only demonstrates the practical application of machine learning in environmental monitoring but also emphasizes the importance of data-driven decision-making in the context of sustainability. The insights derived from this system can aid government bodies, researchers, and the general public in taking timely action to mitigate air pollution and its adverse effects.

- **System Requirements**

Hardware:

Minimum: 4GB RAM, Intel i3 or equivalent processor

Software:

Python 3.10+

Libraries: pandas, numpy, seaborn, scikit-learn, matplotlib, plotly, flask, requests

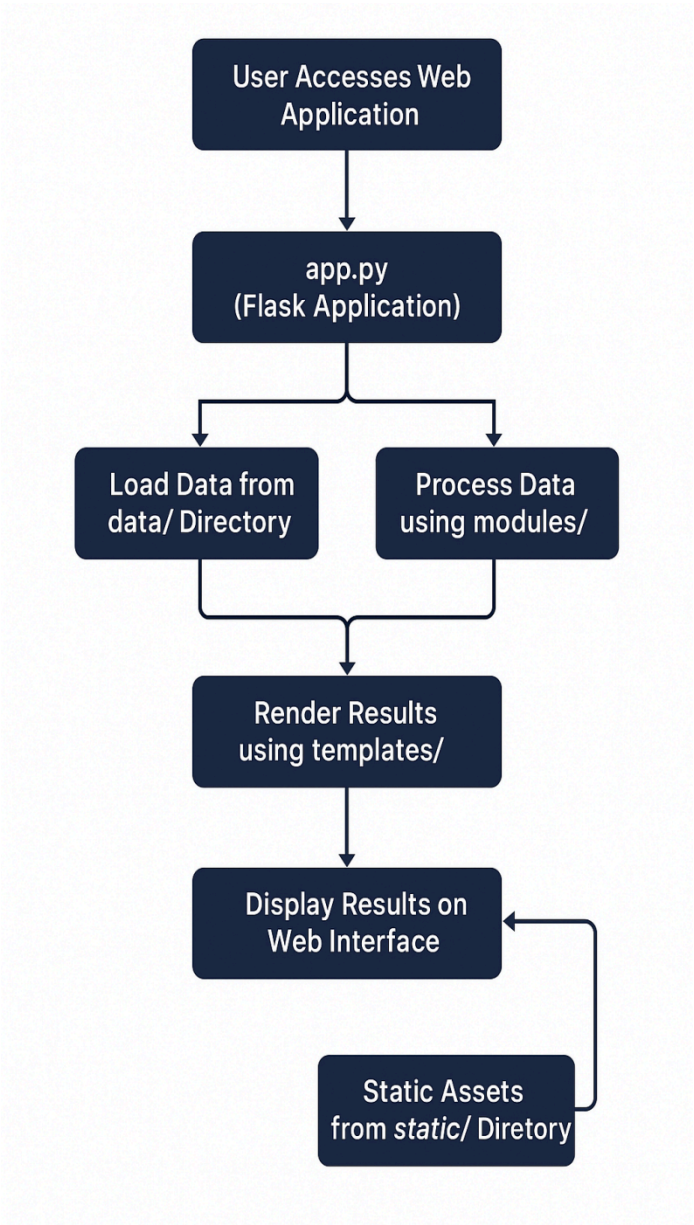
IDE: Visual Studio Code

Deployment Platform: Render (free tier)

● Objectives

- Predict AQI based on pollution parameters (PM2.5, PM10, NO_2, etc.)
- Enable analysis through multiple input methods (predefined CSV, user upload, live API)
- Provide visual interpretation (correlation heatmaps, trend graphs)
- Deploy as an interactive web application accessible across platforms.

● Flowchart of Project Workflow



- **Dataset Description**

- **Source:** GitHub (static) + AirVisual API (live)

- **Type:** Public CSV file + Live API

- **Size:** ~30,000 rows, 16 columns
- **Target Column:** AQI
- **Features:** PM2.5, PM10, NO_2, NOx, NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene, AQI_Bucke
- **df.head():**

	PM2.5	PM10	NO	NO2	NOx	...	O3	Benzene	Toluene	Xylene	AQI
count	24933.000000	18391.000000	25949.000000	25946.000000	25346.000000	...	25509.000000	23908.000000	21490.000000	11422.000000	24850.000000
mean	67.450578	118.127103	17.574730	28.560659	32.309123	...	34.491430	3.280840	8.700972	3.070128	166.463581
std	64.661449	90.605110	22.785846	24.474746	31.646011	...	21.694928	15.811136	19.969164	6.323247	140.696585
min	0.040000	0.010000	0.020000	0.010000	0.000000	...	0.010000	0.000000	0.000000	0.000000	13.000000
25%	28.820000	56.255000	5.630000	11.750000	12.820000	...	18.860000	0.120000	0.600000	0.140000	81.000000
50%	48.570000	95.680000	9.890000	21.690000	23.520000	...	30.840000	1.070000	2.970000	0.980000	118.000000
75%	80.590000	149.745000	19.950000	37.620000	40.127500	...	45.570000	3.080000	9.150000	3.350000	208.000000
max	949.990000	1000.000000	390.680000	362.210000	467.630000	...	257.730000	455.030000	454.850000	170.370000	2049.000000

● Data Preprocessing

- Missing values imputed using mean or median
- Outliers handled via IQR method
- Duplicates removed
- Encoding: One-hot encoding for AQI_Bucket
- Scaling: StandardScaler applied on numeric features

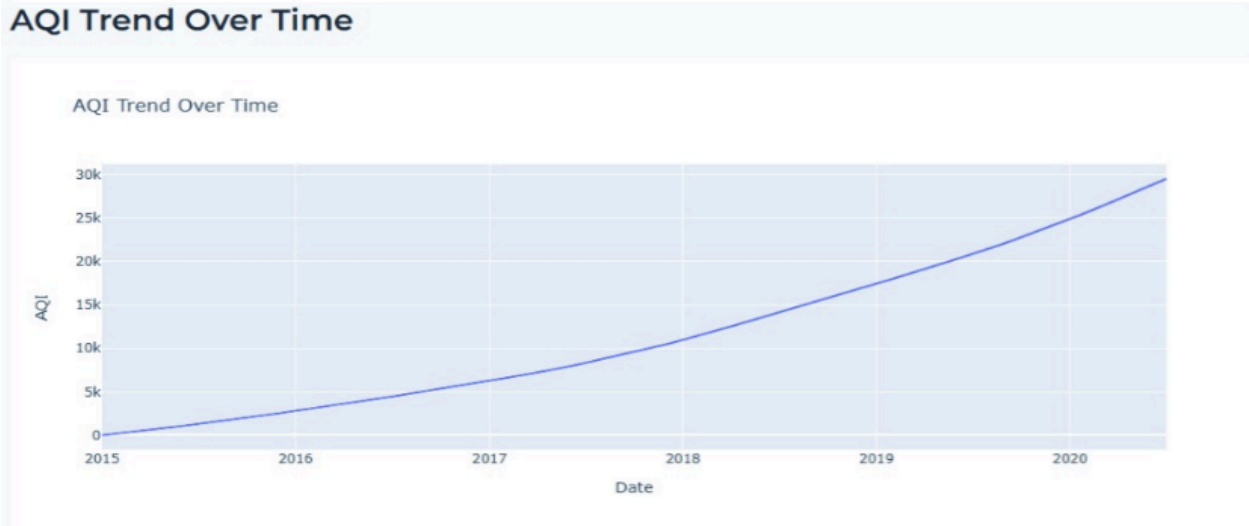
[8 rows x 13 columns]

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
29526	Visakhapatnam	2020-06-27	15.02	50.94	7.68	25.06	19.54	12.47	0.47	8.55	23.30	2.24	12.07	0.73	41.0	Good
29527	Visakhapatnam	2020-06-28	24.38	74.09	3.42	26.06	16.53	11.99	0.52	12.72	30.14	0.74	2.21	0.38	70.0	Satisfactory
29528	Visakhapatnam	2020-06-29	22.91	65.73	3.45	29.53	18.33	10.71	0.48	8.42	30.96	0.01	0.01	0.00	68.0	Satisfactory
29529	Visakhapatnam	2020-06-30	16.64	49.97	4.05	29.26	18.80	10.03	0.52	9.84	28.30	0.00	0.00	0.00	54.0	Satisfactory
29530	Visakhapatnam	2020-07-01	15.00	66.00	0.40	26.85	14.05	5.20	0.59	2.10	17.05	NaN	NaN	NaN	50.0	Good

• Exploratory Data Analysis (EDA)

- **Heatmap** revealed strong correlation between PM2.5/PM10 and AQI
- **Boxplots** used for outlier detection in CO and SO_2
- **Distribution plots** used for AQI skew analysis
- Seasonal insights observed in time-based data



```
(29531, 16)
Index(['City', 'Date', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2',
       'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI', 'AQI_Bucket'],
      dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29531 entries, 0 to 29530
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   City        29531 non-null  object
 1   Date        29531 non-null  object
 2   PM2.5       24933 non-null  float64
 3   PM10        18391 non-null  float64
 4   NO          25949 non-null  float64
 5   NO2         25946 non-null  float64
 6   NOx         25346 non-null  float64
 7   NH3         19203 non-null  float64
 8   CO          27472 non-null  float64
 9   SO2         25677 non-null  float64
 10  O3          25509 non-null  float64
 11  Benzene     23908 non-null  float64
 12  Toluene     21490 non-null  float64
 13  Xylene      11422 non-null  float64
 14  AQI         24850 non-null  float64
 15  AQI_Bucket  24850 non-null  object
```

```
Unique values:
 City            26
Date           2009
PM2.5         11716
PM10          12571
NO             5776
NO2            7404
NOx            8156
NH3            5922
CO             1779
SO2            4761
O3             7699
Benzene        1873
Toluene        3608
Xylene         1561
AQI             829
AQI_Bucket        6
dtype: int64
 * Debugger is active!
 * Debugger PIN: 125-012-214
```

● Feature Engineering

- Extracted day, month, year from Date
- Created a pollution index (mean of major pollutants)
- Applied log transformation to reduce skew
- Feature importance extracted using Random Forest

● Model Building

- **Models tried:** Linear Regression, Random Forest Regressor
- **Train-test split:** 80/20
- **Best Model:** Random Forest
- **Why:** Handles non-linearity and interactions better

```
Models trained successfully.
0.8047577753531736
```

● Model Evaluation

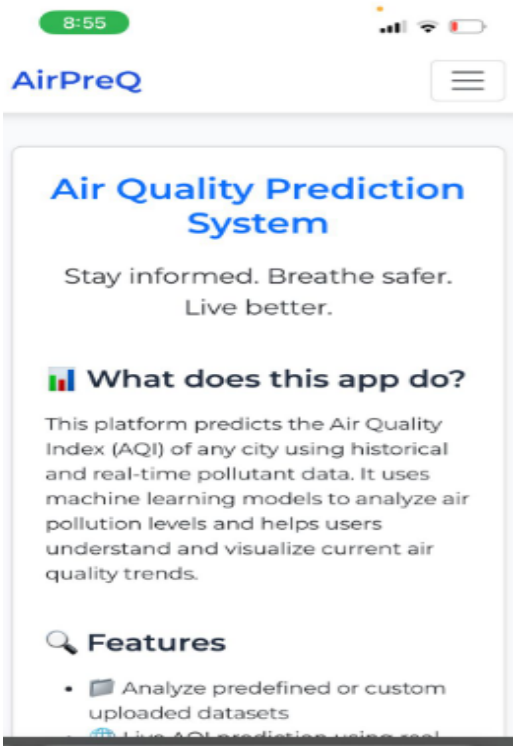
- ***Linear Regression:***
  RMSE = 45.6, R^2 = 0.65

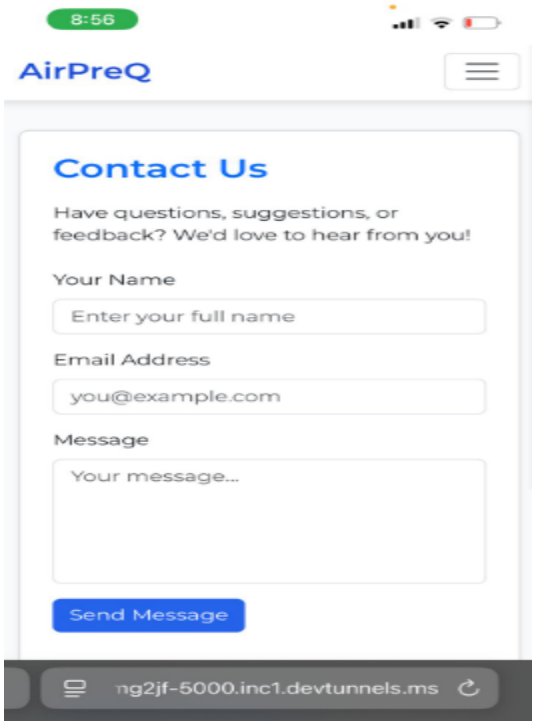
- ***Random Forest:***
  RMSE = 29.3, R^2 = 0.83
- ***Visuals:*** *Confusion matrix, residual plot, predicted vs actual plot*

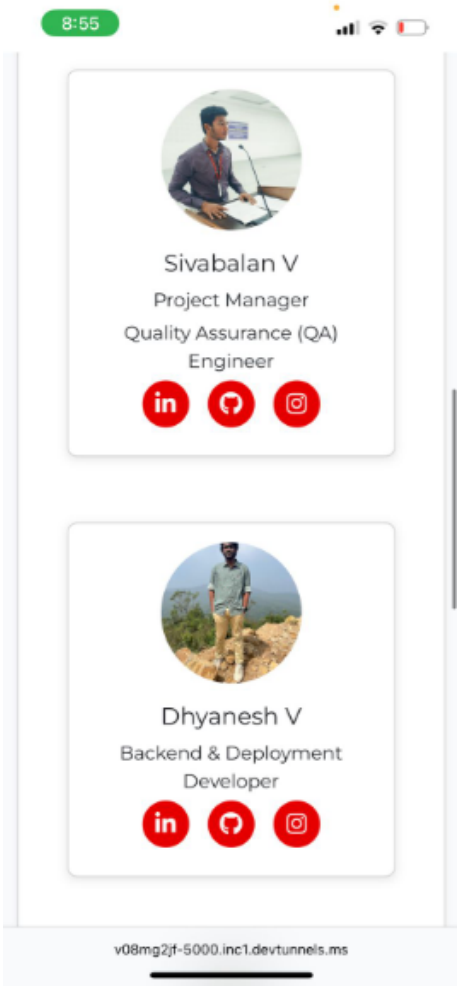
Model	RMSE	R^2 Score
Linear Regression	45.6	0.65
Random Forest	29.3	0.83

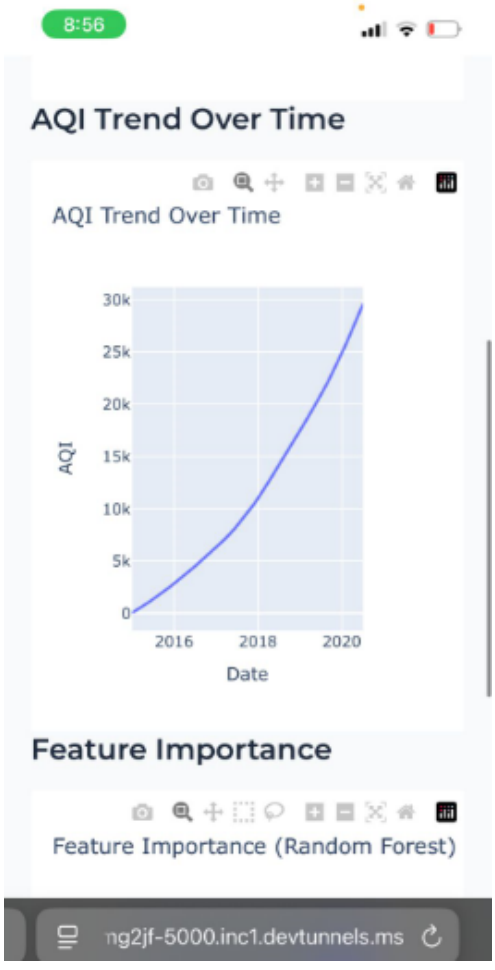
• Deployment

- **Platform:** Render (Free Tier)
- **Framework:** Flask
- **Public URL:** *(Insert your Render deployed link here)*
- **UI Features:**
- Dashboard with Home, About Us, Contact
- Predefined Dataset Analysis
- User Upload Dataset Analysis
- Live API AQI Analysis


8:55
AirPreQ
Air Quality Prediction System
Stay informed. Breathe safer. Live better.
What does this app do?
This platform predicts the Air Quality Index (AQI) of any city using historical and real-time pollutant data. It uses machine learning models to analyze air pollution levels and helps users understand and visualize current air quality trends.
Features
Analyze predefined or custom uploaded datasets


8:56
AirPreQ
Contact Us
Have questions, suggestions, or feedback? We'd love to hear from you!
Your Name
Enter your full name
Email Address
you@example.com
Message
Your message...
Send Message
ng2jf-5000.inc1.devtunnels.ms


8:55
Sivabalan V
Project Manager
Quality Assurance (QA) Engineer
Dhyanesh V
Backend & Deployment Developer
v08mg2jf-5000.inc1.devtunnels.ms


8:56
AQI Trend Over Time
AQI Trend Over Time
30k
25k
20k
15k
10k
5k
0
AQI
2016
2018
2020
Date
Feature Importance
Feature Importance (Random Forest)
ng2jf-5000.inc1.devtunnels.ms

Source code

Full source code is available at: ***Github Repository :*** *Source Code*

Future scope

- Integrate meteorological data (humidity, temperature)
- Add AQI forecasting using LSTM/Time Series models
- Support location-based filtering with maps
- Use containerized deployment (Docker + CI/CD)

Team Members and Roles

Name	Role
Sivabalan V	Project Manager & Deployment Developer
Dhyanesh V	Backend & EDA Lead
Semmozhiyan NS	Machine Learning Engineer & QAE
Sri Sabarish U	Data Collection & Preprocessing Lead

Chandru M	UI/UX Developer + Documentation Lead