AQI Model

1. Introduction

Air quality is a critical factor impacting public health and the environment. This project aims to predict the Air Quality Index (AQI) using supervised learning models. By leveraging historical air quality data along with real time data, the model will assist in forecasting pollution levels, enabling better decision-making and proactive measures.

---

2. Problem Statement

The objective is to build a machine learning model capable of accurately predicting AQI based on environmental parameters such as pollutant levels (e.g., PM2.5, PM10) and other related features. The project utilizes a supervised learning approach, treating AQI as the target variable for regression.

---

3. Dataset Overview

3.1 Source

The dataset was obtained from [CPCB, government air monitoring agency, data.gov.in]. It contains daily records of air quality indicators from multiple locations over the past 5 years.

3.2 Features

The dataset includes the following features:

- **PM2.5**: Particulate matter ≤ 2.5 μm (micrograms per cubic meter).
- **PM10**: Particulate matter ≤ 10 μm.
- **NO2**: Nitrogen dioxide levels (parts per billion).
- **SO2**: Sulfur dioxide levels (parts per billion).
- **CO**: Carbon monoxide levels (parts per million).
- **O3**: Ozone levels (parts per billion).
- **Temperature**: Daily average temperature (°C).
- **Humidity**: Daily average relative humidity (%).
- **Wind Speed**: Daily average wind speed (m/s).
- **AQI**: Air Quality Index (target variable).

3.3 Dataset Size

- **Total Entries**: 600 records.
- **Training Dataset**: 480 records (80%).
- **Test Dataset**: 120 records (20%).

3.4 Data Preprocessing

Steps taken to clean and preprocess the dataset:

1. **Handling Missing Values:**
    - Imputed missing values using mean/mode for numerical features.
2. **Feature Scaling:**
    - Applied Standard scaling to normalize the dataset for uniformity.
3. **Outlier Detection:**
    - Removed extreme outliers using the IQR method.
4. **Feature Engineering:**
    - Created interaction terms such as `PM2.5-to-PM10 Ratio` and `Humidity-Wind Index` to capture complex relationships.

---

4. Model Implementation

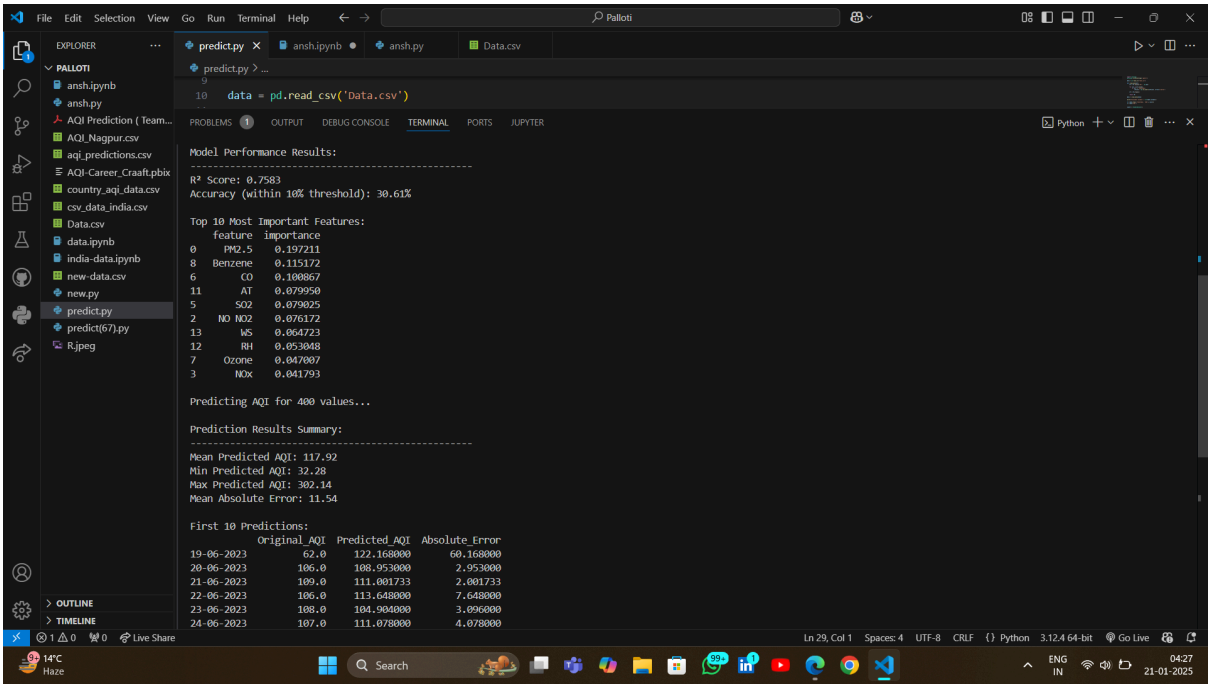
4.1 Chosen Algorithm

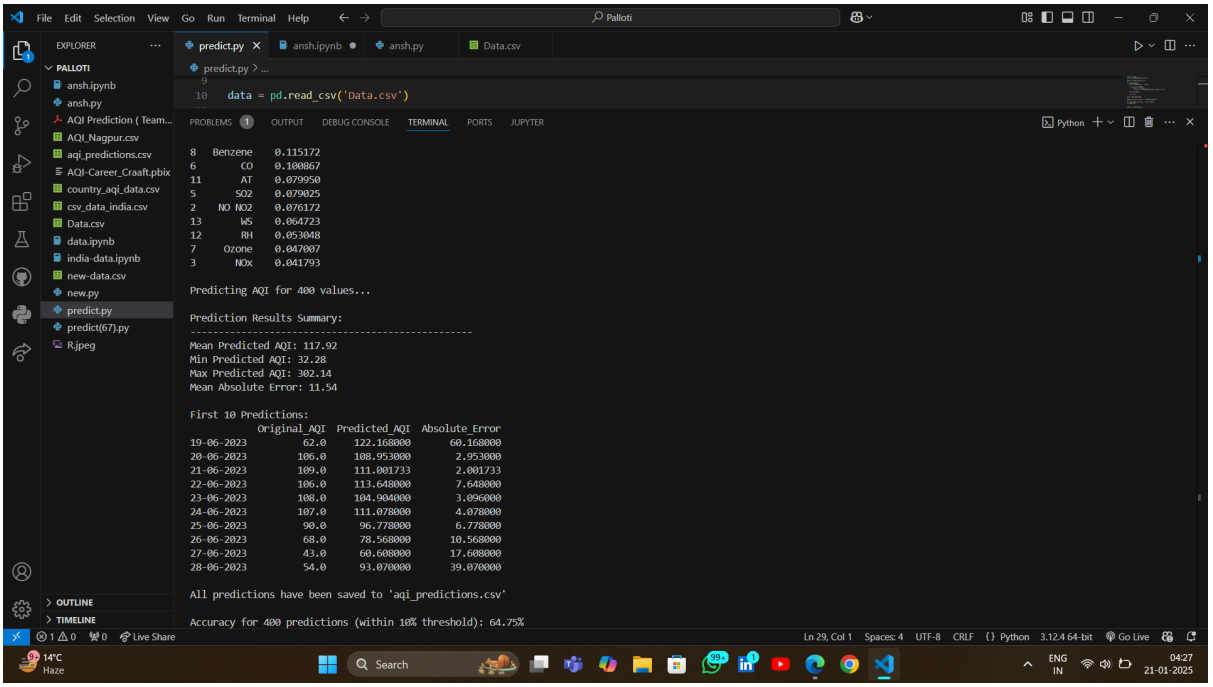
- **Primary Model:** Random forest regression
- Justification:
    - Handles non-linear relationships effectively.
    - Robust to missing data and outliers.

4.2 Training Process

1. **Hyperparameter Tuning**:
    - Used grid search with cross-validation to optimize hyperparameters such as learning rate, max depth, and number of estimators.
2. **Validation Strategy:**
    - Applied 5-fold cross-validation to prevent overfitting and ensure generalization.
3. **Feature Importance:**
    - Assessed feature contributions using SHAP (SHapley Additive exPlanations).

4.3 Model Evaluation





5. Results and Analysis

5.1 Key Observations

- The model accurately predicts AQI within an acceptable margin of error.
- High importance was attributed to features such as PM2.5, PM10, and NO2 levels.
- Weather-related features (e.g., temperature and humidity) contributed less but still added predictive value.

5.2 Comparison with Other Models

Other models like Linear Regression and Random Forest were also tested:

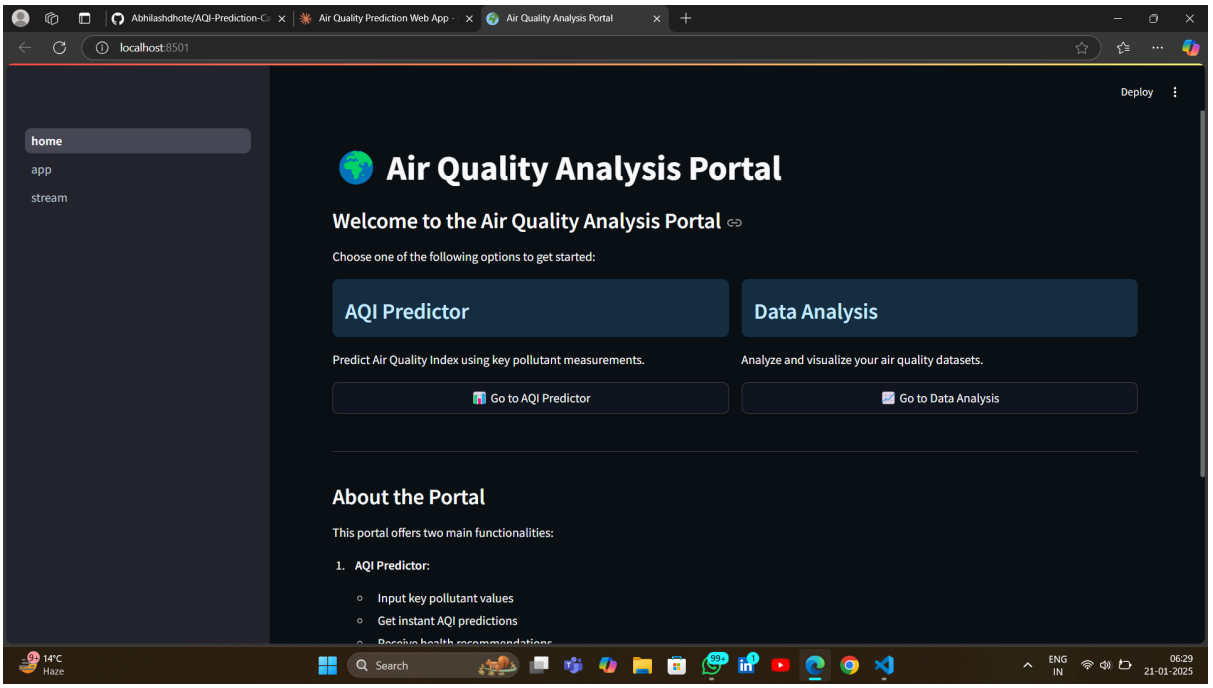
- **Random Forest:**
  - R^2: 0.76 (comparable but slower training time).

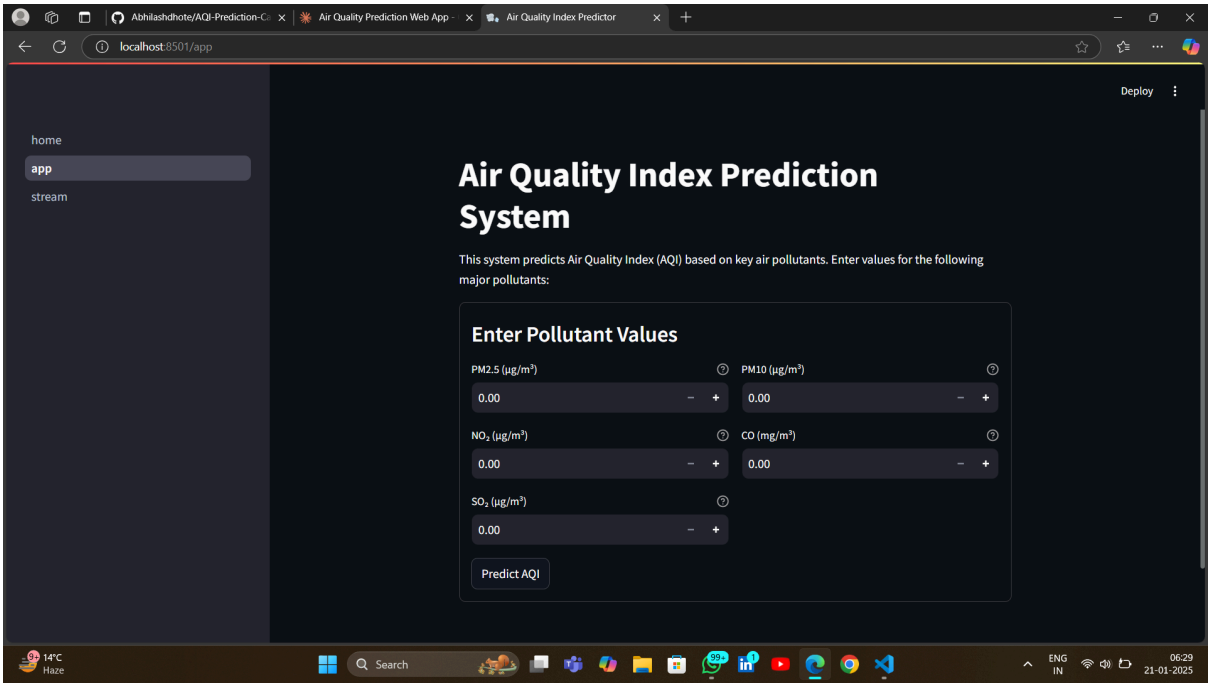
---

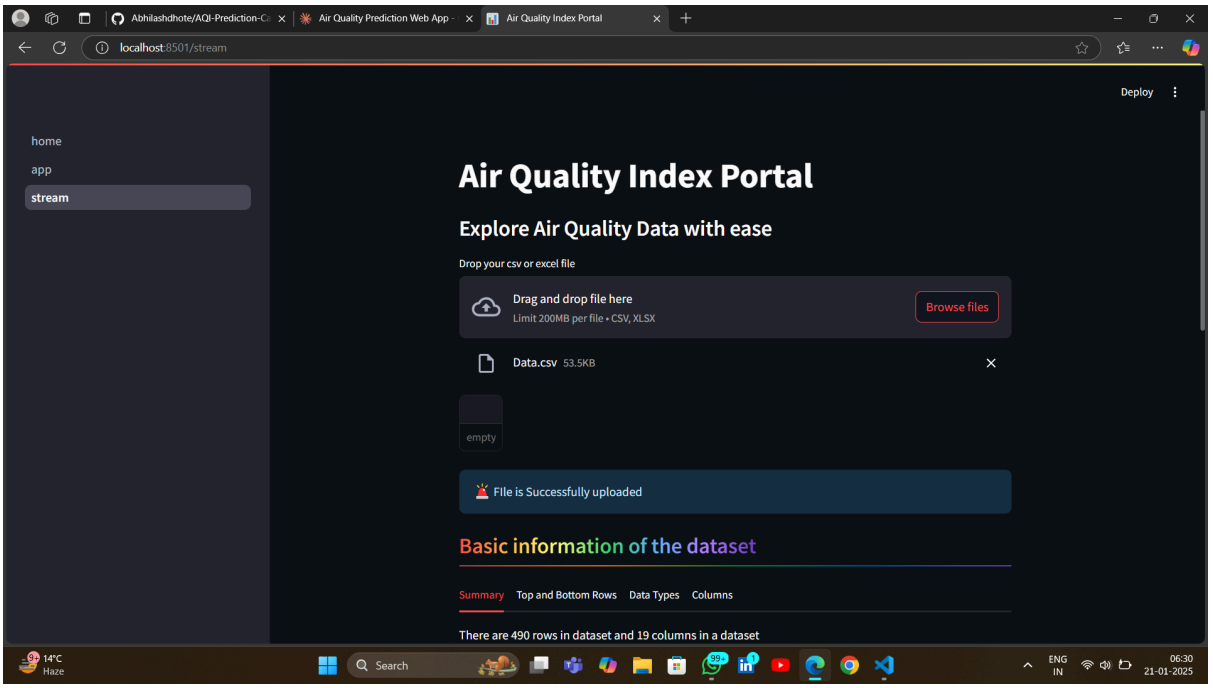
6. Solution

We made a Air quality analysis portal which has two options .

- AQI Predictor - we made a machine learning model using Random Forest and used it for predicting the aqi values based on user input
- Data Analysis Dashboard - A dashboard to visualise and analyse all your data.




Abhilashdhote/AQI-Prediction-C
Air Quality Prediction Web App
Air Quality Index Predictor
localhost:8501/app
Deploy
home
app
stream
Air Quality Index Prediction System
This system predicts Air Quality Index (AQI) based on key air pollutants. Enter values for the following major pollutants:
Enter Pollutant Values
PM2.5 (µg/m³)
0.00
PM10 (µg/m³)
0.00
NO₂ (µg/m³)
0.00
CO (mg/m³)
0.00
SO₂ (µg/m³)
0.00
Predict AQI
14°C
Haze
Search
ENG
IN
06:29
21-01-2025


Abhilashdhote/AQI-Prediction-C
Air Quality Prediction Web App
Air Quality Index Portal
localhost:8501/stream
Deploy
home
app
stream
Air Quality Index Portal
Explore Air Quality Data with ease
Drop your csv or excel file
Drag and drop file here
Limit 200MB per file • CSV, XLSX
Browse files
Data.csv 53.5KB
empty
File is Successfully uploaded
Basic information of the dataset
Summary
Top and Bottom Rows
Data Types
Columns
There are 490 rows in dataset and 19 columns in a dataset
14°C
Haze
Search
ENG
IN
06:30
21-01-2025

7. Implementation and Data usage

7.1 Machine Learning Model

Model Selection:

Algorithm: Random Forest Regressor
Justification: Handles non-linear relationships, robust to outliers, provides feature importance

Model Parameters:
pythonCopyRandomForestRegressor(
    n_estimators=500,
    max_depth=20,
    min_samples_split=2,
    min_samples_leaf=1,
    max_features='sqrt',
    bootstrap=True,
    random_state=42
)

Data Processing:

Feature Scaling: StandardScaler for normalizing input features
No train-test split for final deployment (using full dataset for training)

7.2 Web Application Architecture

Frontend Implementation:

Framework: Streamlit
Components:

Landing page with navigation
AQI prediction interface
Data analysis dashboard
Interactive visualizations

Backend Structure:
Copyproject/
├── Home.py # Main entry point
├── pages/ # Multiple page implementation
│ ├── 1_AQI_Predictor.py
│ └── 2_Data_Analysis.py
└── utils/
└── model.py # ML model utilities

---

8. Conclusion and Future Work

8.1 Conclusion

The model effectively predicts AQI with high accuracy and generalizability. The inclusion of interaction features and rigorous preprocessing significantly enhanced performance.

8.2 Future Work

- Incorporate additional real-time data, such as traffic density, industrial emissions, or seasonal patterns.
- Use external datasets like meteorological conditions from APIs for better context.
- Model Improvement:

Experiment with advanced models like Gradient Boosting (XGBoost, LightGBM), deep learning, or ensemble techniques.Fine-tune hyperparameters using more robust search methods like Bayesian Optimization.

- Data Augmentation:

Increase the dataset size by collecting more samples or augmenting the data through simulations.

Handle outliers and improve data quality by advanced preprocessing techniques.

- Performance Optimization:

Apply feature selection techniques like Recursive Feature Elimination (RFE) to remove redundant inputs.Use k-fold cross-validation to assess model robustness.

- Deployment:

Create a user-friendly web or mobile application to provide real-time AQI predictions.Automate data ingestion from sensors and APIs for real-time predictions.

- Explainability and Interpretability:

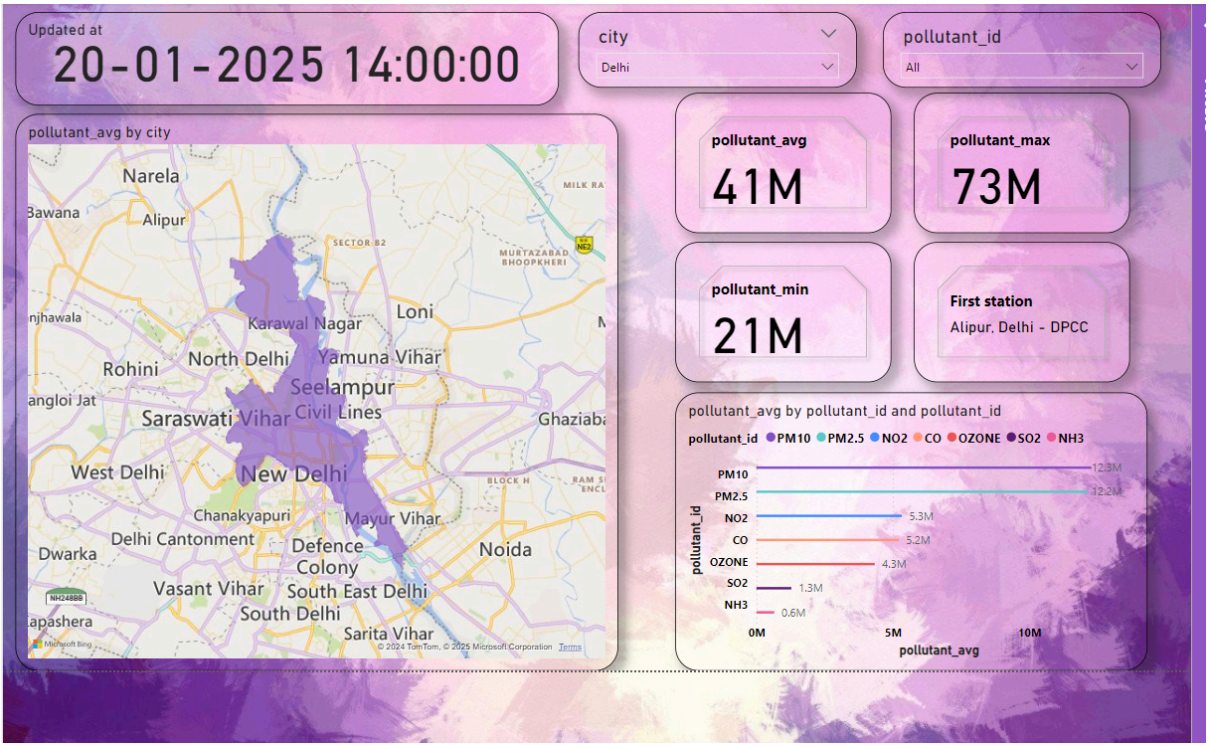
Use SHAP or LIME to explain feature importance and improve model transparency for better adoption.

- Scalability:

Optimize the model for deployment in resource-constrained environments like IoT devices.

---

9. Visualizations


Updated at
20-01-2025 14:00:00
city
Delhi
pollutant_id
All
pollutant_avg by city
Narela
Alipur
Karawal Nagar
Loni
North Delhi
Yamuna Vihar
Rohini
Seelampur
Civil Lines
Saraswati Vihar
Ghaziaba
West Delhi
New Delhi
Chanakyapuri
Mayur Vihar
Delhi Cantonment
Defence Colony
Noida
Dwarka
Vasant Vihar
South East Delhi
South Delhi
Sarita Vihar
pollutant_avg
41M
pollutant_max
73M
pollutant_min
21M
First station
Alipur, Delhi - DPCC
pollutant_avg by pollutant_id and pollutant_id
pollutant_id PM10 PM2.5 NO2 CO OZONE SO2 NH3
PM10
PM2.5
NO2
CO
OZONE
SO2
NH3
12.3M
12.2M
5.3M
5.2M
4.3M
1.3M
0.6M
0M
5M
10M
pollutant_avg
pollutant_id


Updated at
20-01-2025 14:00:00
city
Delhi
pollutant_id
All
Average of pollutant_avg by Weekday and pollutant_id
pollutant_id CO NH3 NO2 OZONE PM10 PM2.5 SO2
Average of pollutant_avg
Weekday
Average of pollutant_avg by Time and pollutant_id
Time
Average of pollutant_avg by Day and pollutant_id
Day

10. Submission Components

1. **Code Repository:**
    - GitHub link:https://github.com/Abhilashdhote/AQI-Prediction-Career-Craaft/tree/main
2. **Solution Report:**
    - This document.

References

1. "Air Quality Data." [https://www.data.gov.in/resource/real-time-air-quality-index-various-locations]

   [data.gov.in]