AQI Prediction Report

The project were to collect and clean weather and pollutant data, perform exploratory data analysis (EDA), and study relationships between environmental conditions and AQI levels. After understanding the dataset, several machine learning models were trained and evaluated to determine the most accurate predictor of AQI.

Dataset Description

The dataset consists of weather and pollutant variables such as temperature (°C), humidity (%), PM2.5, PM10, NO_2, CO, and the AQI values. AQI was calculated using pollutant concentrations, which reflect air quality ranging from “Good” to “Hazardous.” The dataset is structured, cleaned, and prepared for statistical and machine learning analysis.

Exploratory Data Analysis

The EDA revealed strong seasonal and environmental influences on AQI. Higher PM2.5 levels were consistently linked with worse AQI values, while high humidity reduced pollutant concentrations, improving air quality. AQI remained relatively stable within the “Unhealthy” range, with sudden spikes caused by particulate matter.

Statistical Summary

The dataset showed a mean AQI of 170, with a minimum of 154 and a maximum of 190. A standard deviation of 12.5 indicates a fairly stable but consistently unhealthy level of air pollution.

Machine Learning Models

Multiple models were trained to predict AQI, including Decision Tree, Logistic Regression, Support Vector Machine, KNN, and Random Forest. Among these, Random Forest provided the highest accuracy (89%) due to its ability to handle complex, non-linear data patterns. Logistic Regression and SVM performed moderately, while KNN achieved the lowest accuracy.

Machine Learning Models & Accuracy

Several models were trained and compared for AQI prediction:

- Random Forest → 89% (Best Performance)
- Decision Tree → 84%
- Support Vector Machine (SVM) → 76%
- Logistic Regression → 74%

- K-Nearest Neighbors (KNN) → 71%

Code Implementation

The code involved several key steps:

1. Data preprocessing, including handling missing values and scaling features.
2. Data visualization using Python libraries such as Matplotlib and Seaborn.
3. Model training with RandomForestRegressor and other algorithms.
4. Model evaluation using metrics such as R^2 Score and Mean Squared Error.
5. Prediction of AQI for unseen weather and pollutant data.

Prediction Output

```
Prediction for the next 3 time records (simulated 3-day prediction):

Timestamp: 2025-08-17 00:13:06
RandomForest        : Predicted AQI Index (1-5): 2
DecisionTree        : Predicted AQI Index (1-5): 2
LogisticRegression  : Predicted AQI Index (1-5): 2
KNeighbors          : Predicted AQI Index (1-5): 2
SVC                 : Predicted AQI Index (1-5): 2

Timestamp: 2025-08-17 00:14:07
RandomForest        : Predicted AQI Index (1-5): 2
DecisionTree        : Predicted AQI Index (1-5): 2
LogisticRegression  : Predicted AQI Index (1-5): 2
KNeighbors          : Predicted AQI Index (1-5): 2
SVC                 : Predicted AQI Index (1-5): 2

Timestamp: 2025-08-17 00:15:08
RandomForest        : Predicted AQI Index (1-5): 2
DecisionTree        : Predicted AQI Index (1-5): 2
LogisticRegression  : Predicted AQI Index (1-5): 2
KNeighbors          : Predicted AQI Index (1-5): 2
SVC                 : Predicted AQI Index (1-5): 2
```

The Random Forest model predicted an AQI of 182, This demonstrates how the model can provide real-time forecasts for public awareness

Conclusion

The project concludes that particulate matter (PM2.5 and PM10) are the most critical pollutants influencing AQI. Humidity acts as a natural purifier by lowering pollutant concentrations. The Random Forest model was found to be the most effective, achieving 89% accuracy. This system

can provide valuable insights for government agencies, environmental organizations, and citizens to take preventive actions against air pollution.