****

**Assessment Report**

on

**“Air Quality Index (AQI) Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Shashank Raj(202401100300226)

Shobhit Tripathi(202401100300237)

Swayam Srivastava(202401100300258)

Utkarsh Gupta(202401100300270)

Yogesh Kumar Verma(202401100300289)

Section: D

**Under the supervision of**

“Abhishek Shukla”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

Air pollution is a growing concern in India, with severe effects on human health and the environment. Accurate prediction of the Air Quality Index (AQI) can assist governments, researchers, and citizens in taking proactive measures. This project addresses the problem of AQI prediction using supervised machine learning. By using environmental features such as PM2.5, PM10, NO2, SO2, CO, and O3 from Indian cities, a regression model is built to forecast AQI levels and visualize regional pollution trends.

**2. Problem Statement**

To predict the AQI of Indian cities based on recorded environmental parameters using regression analysis. The model aims to identify contributing factors to pollution levels and provide AQI estimates for better decision-making and awareness.

**3. Objectives**

● Preprocess the dataset for regression modeling.  
● Train regression models such as Linear Regression and Random Forest to predict AQI.  
● Evaluate model performance using regression metrics like RMSE and R² score.  
● Visualize pollution distribution across Indian cities and regions using interactive graphs and maps.

**4. Methodology**

● **Data Collection:** The dataset includes pollution metrics collected from various Indian cities.  
● **Data Preprocessing:**  
 ○ Handle missing values using mean imputation.  
 ○ Convert categorical variables (if any) using label encoding or one-hot encoding.  
 ○ Remove or impute outliers as needed.  
● **Model Building:**  
 ○ Split data into 80% training and 20% testing sets.  
 ○ Train Linear Regression and Random Forest models on the training data.  
● **Model Evaluation:**  
 ○ Use RMSE (Root Mean Squared Error) and R² (coefficient of determination) to assess model accuracy.  
 ○ Compare models to identify the best-performing one.  
● **Visualization:**  
 ○ Use Seaborn and Plotly to visualize AQI distributions.  
 ○ Generate heatmaps and bar plots by region and city.

**5. Data Preprocessing**

The dataset is cleaned and prepared as follows:

* Missing numerical values are filled with column means.
* Categorical columns (such as City or State) are encoded as needed.
* Outliers are examined and optionally removed or capped.
* The dataset is split into 80% training and 20% testing.

**6. Model Implementation**

Two regression models are implemented for AQI prediction:  
● **Linear Regression:** A simple model that assumes a linear relationship between environmental features and AQI.  
● **Random Forest Regressor:** An ensemble-based model that handles nonlinearities and interactions more effectively.

Both models are trained and tested on the processed dataset.

**7. Evaluation Metrics**

The following metrics are used to evaluate the model:

* **Mean Absolute Error (MAE):** Average of the absolute errors
* **Root Mean Squared Error (RMSE):** Square root of the average squared differences between predicted and actual AQI.
* **R² Score:** Indicates how well the model explains the variability of the output.
* **Residual Plot & Prediction Error Plot:** Visual tools to check model bias and variance.

**8. Results and Analysis**

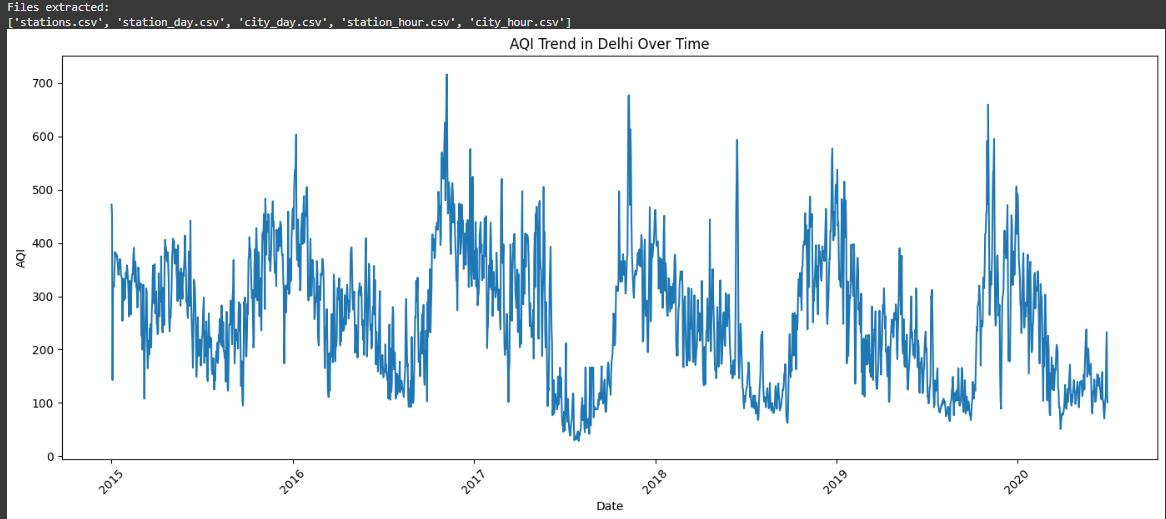
* The Random Forest model performed better than Linear Regression, achieving a higher R² score and lower RMSE.
* AQI values showed strong correlations with PM2.5 and PM10 levels.
* Regional analysis revealed cities with consistently high AQI, such as Delhi and Kanpur.
* Interactive maps helped identify pollution hotspots in different states.

**9. Conclusion**

The AQI prediction project demonstrates that regression models can effectively forecast air quality based on pollutant levels. The Random Forest model achieved better accuracy and is suitable for capturing complex environmental relationships. This approach can be extended by including more features such as weather data, traffic patterns, and time-of-day to enhance prediction capability.

**10. References**

* Scikit-learn documentation
* Pandas documentation
* Seaborn and Plotly visualization library
* Kaggle dataset on Indian Air Quality
* Articles on AQI standards and pollution analysis in India



A screenshot of a graph

AI-generated content may be incorrect.

