KIET Group of insitution
Department of Computer Science & Engineering (AI)

Air Quality Data in india

Air Quality Index (AQI) and hourly data across stations and cities in India

group :-3

Abhishek Kumar Maurya
Aryan Raj Pandey
Ansh Tandon
Aditya Singh
Anurag

2015-2020

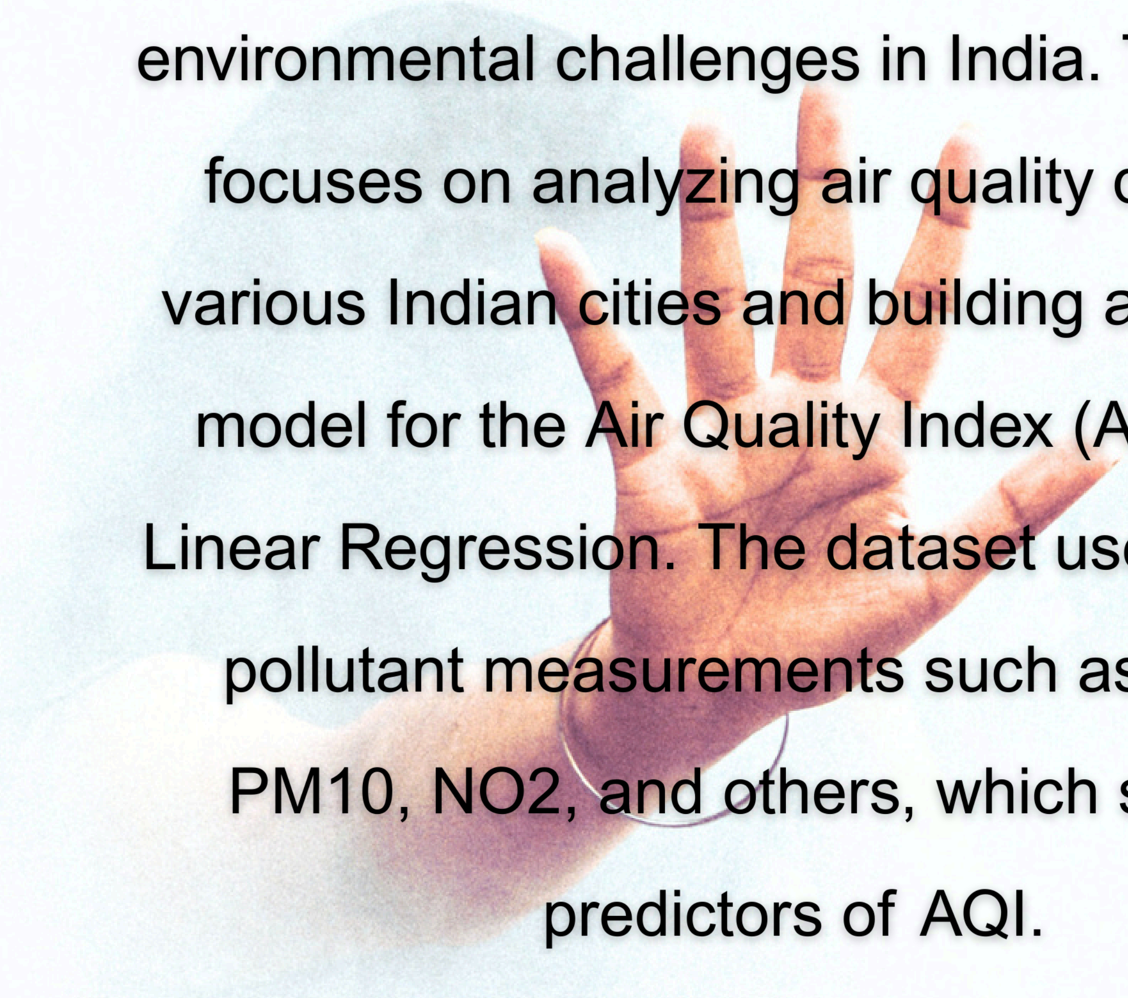
Submitted By:

[GROUP-3]
B.Tech CSE (AI)

Submitted To:

[BIKKI kumar]
Department of Computer Science & Engineering (AI)

introduction



Air pollution poses serious health and environmental challenges in India. This project focuses on analyzing air quality data from various Indian cities and building a predictive model for the Air Quality Index (AQI) using Linear Regression. The dataset used includes pollutant measurements such as PM2.5, PM10, NO2, and others, which serve as predictors of AQI.

PROBLEM STATEMENT

To analyze historical air quality data in India and develop a machine learning model to predict AQI levels. This will help in identifying highly polluted regions and enable early warnings for public health safety.

OBJECTIVES

- Load and explore the air quality dataset (city_day.csv).
- Handle missing values and prepare clean data for modeling.
- Visualize the relationship between AQI and major pollutants.
- Build a Linear Regression model to predict AQI.
- Evaluate the model using regression metrics.

METHODOLOGY

● Data Collection

- The dataset city_day.csv was loaded using pandas.
- It contains pollutant readings and AQI data across Indian cities.

● Exploratory Data Analysis (EDA)

- Basic shape and structure of the dataset examined.
- Missing values per column identified.
- Key relationships visualized:
  - Distribution of AQI.
  - Scatter plots: AQI vs PM2.5, AQI vs PM10.
  - Box plot: AQI distribution across selected cities.

● Data Cleaning

- Numerical columns with missing values (e.g., PM2.5, NO2, etc.) were imputed using median.
- Categorical column AQI_Bucket was filled using mode.
- Verified that all missing values were successfully handled.

● Feature Selection

- Target variable: AQI
- Input features: All numerical pollutant columns except AQI (e.g., PM2.5, PM10, NO, NO2, CO, etc.)

● Model Training

- The data was split into training (80%) and testing (20%) using train_test_split.
- A Linear Regression model was trained on the features to predict AQI.

● Model Evaluation

- Predictions were made on the test set.
- Performance metrics computed:
  - Mean Squared Error (MSE)
  - R-squared (R^2 Score)

RESULTS AND ANALYSIS

✅ **Visual Insights**:

- AQI Distribution showed a skew towards moderate to poor air quality.
- Strong positive correlation was observed between AQI and PM2.5 / PM10.
- AQI varied significantly across cities (shown via box plot).

📈 **Model Performance**:

- Mean Squared Error: [insert output value here]
- R-squared Score: [insert output value here]

These metrics indicate how well the model fits the data. A higher R^2 score implies a better fit.

📌 **Observation:**

- Linear Regression performed reasonably well as a baseline model.
- PM2.5 and PM10 emerged as dominant predictors of AQI.

CONCLUSION

The project successfully implemented a basic AQI prediction model using Linear Regression. The analysis revealed strong pollutant-AQI relationships, especially with PM2.5 and PM10. While Linear Regression provides a simple starting point, more complex models (like Random Forest, XGBoost) and time-series methods could improve prediction accuracy.

FUTURE SCOPE

- Incorporate temporal data (e.g., seasons, months) to capture seasonal pollution trends.
- Use classification models to predict AQI categories (Good, Moderate, Poor, etc.).
- Address class imbalance and test more robust models like Random Forest, SVR, or XGBoost.
- Build real-time AQI dashboards for public use.

REFERENCES

- scikit-learn documentation – https://scikit-learn.org
- pandas documentation – https://pandas.pydata.org
- Seaborn for visualizations – https://seaborn.pydata.org

CPCB India – https://cpcb.nic.in