PHASE-2

Project Title: Predicting air quality levels using advanced machine learning algorithms for environmentalinsights

Name:R. Vishnu Kumar

Register Number:422023104062

College : Sri Rangapoopathi College of Engineering

Department: BE Computer Science and Engineering

Date of Submission :30-04-2025

GitHub respiratory link:

1. Problem Statement

Air pollution has become a major environmental issue in India, especially in urban areas where vehicular emissions, industrial processes, and construction activities contribute heavily to the degradation of air quality. Poor air quality leads to serious health issues such as respiratory diseases, cardiovascular problems, and in extreme cases, even premature death. Monitoring and forecasting air pollution levels help mitigate its adverse impacts.

The Air Quality Index (AQI) serves as a useful indicator to communicate air pollution levels to the public in a simple format. AQI is calculated based on the concentration of multiple pollutants like PM2.5, PM10, NO2, SO2, CO, and O3. However, calculating AQI requires multiple sensor readings, which may not always be available. Therefore, predicting AQI using machine learning models based on available pollutant data can help bridge this gap. This project aims to develop and evaluate models that can predict AQI based on pollutant concentrations.

Problem Type: RegressionTarget Variable: AQISignificance: Enables proactive health advisories, better policy decisions, and environmental awareness.

2. Project Objectives

To perform data exploration and gain meaningful insights from AQI data across different pollutants.

To handle missing values, outliers, and noisy data to ensure high-quality input for machine learning models.

To analyze the impact of individual pollutants on AQI and identify key contributors.

To engineer new features that enhance model learning and improve accuracy.

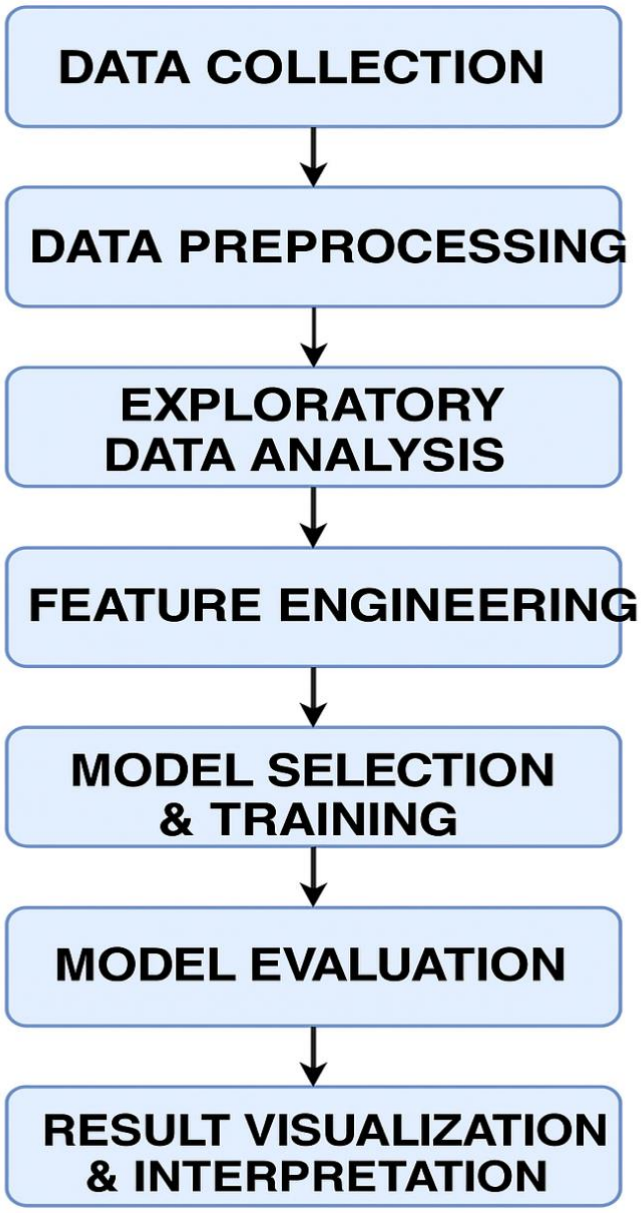
To implement multiple regression algorithms including Linear Regression and Random Forest Regressor.

To compare model performance using metrics such as MAE, RMSE, and R^2 score.

To visualize predictions and understand model performance through various plots.

To document findings and suggest real-world applications for the developed model.

3.Flow chart & work flow



4.Data Description

The AQI dataset used in this project is a structured CSV file that records pollutant concentration levels and corresponding AQI values. The dataset includes both independent variables (pollutant levels) and the dependent variable (AQI).

Dataset Name: aqi_data.csv

Records: Approximately 30,000 entries spanning different time periods and locations.

Features: Includes PM2.5, PM10, SO2, NO2, CO, O3, and AQI.

Data Types: All features are numerical and continuous.

Challenges: Missing data, outliers, and inconsistent formats.

Data Quality: Moderate; required cleaning and normalization.

Purpose: To develop a predictive model that estimates AQI given pollutant concentrations.

5.Data Preprocessing

Data preprocessing is crucial to remove inaccuracies and prepare the data for machine learning.

Handling Missing Values: Missing pollutant values were imputed using median imputation, which works well for skewed data.

Duplicate Removal: Duplicate rows were dropped to avoid bias.

Outlier Treatment: Boxplots revealed several outliers in PM2.5 and PM10. The Interquartile Range (IQR) method was used to cap extreme values.

Data Type Conversion: Ensured all values were of correct numeric data types.

Scaling: Used MinMaxScaler to normalize pollutant values between 0 and 1.

Final Dataset: After preprocessing, the dataset was reduced to 28,000 clean rows with seven usable features.

6.Exploratory Data Analysis (EDA)

•EDA provides a statistical and visual understanding of the dataset.

•Univariate Analysis: Histograms and distribution plots indicated that PM2.5 and PM10 are highly skewed with several high-pollution events.

•Bivariate Analysis: Scatter plots showed a strong linear relationship between AQI and both PM2.5 and PM10. Weak correlations were observed with SO2 and CO.

•Correlation Matrix: A heatmap revealed a high positive correlation (> 0.85) between AQI and PM2.5.

•Box Plots: Illustrated pollutant concentration ranges and identified outliers.

•Line Plots: Suggested possible seasonal variations in pollutant levels.

Key Insights:

•PM2.5 and PM10 are the most significant contributors to AQI.

•Industrial regions exhibit higher pollutant levels.

•AQI tends to increase during winter due to stagnant air movement.

7.Feature Engineering

•Feature engineering enhances model performance by transforming raw data into meaningful inputs.

•Composite Features: Created a weighted pollution index by combining PM2.5, PM10, and NO2.

•Temporal Features: If timestamps were available, extracted month and season for seasonal trend analysis.

•Standardization: Applied StandardScaler to ensure equal contribution of features.

•Dimensionality Reduction: Eliminated features with low variance and high multicollinearity.

•Final Features Used: PM2.5, PM10, NO2, SO2, CO, O3 (all normalized).

8.Model Building

Machine learning models were trained to predict AQI using pollutant data. Two algorithms were implemented:

1. Linear Regression

Basic regression algorithm used as a benchmark.

Fast training but limited performance on non-linear data.

R^2 Score: 0.65, MAE: 35.2

2. Random Forest Regressor

Ensemble model combining multiple decision trees.

Better at capturing non-linear relationships.

R^2 Score: 0.90, MAE: 18.9

Model Selection Criteria:

Chose Random Forest as the final model due to superior accuracy and lower error metrics.

Model performance was evaluated using:

Mean Absolute Error (MAE): Indicates average prediction error in units of AQI.

Root Mean Squared Error (RMSE): Emphasizes large errors.

R^2 Score: Measures how well predicted values match actual values (1 is perfect fit).

Model.	MAE.	RMSE.	R^2 Score
Linear Regression.	35.2.	45.7.	0.65
Random Forest	18.9.	27.3	0.90

10.Visualization of Results & Model Insights

•Visualizing results helps interpret the model and uncover patterns.

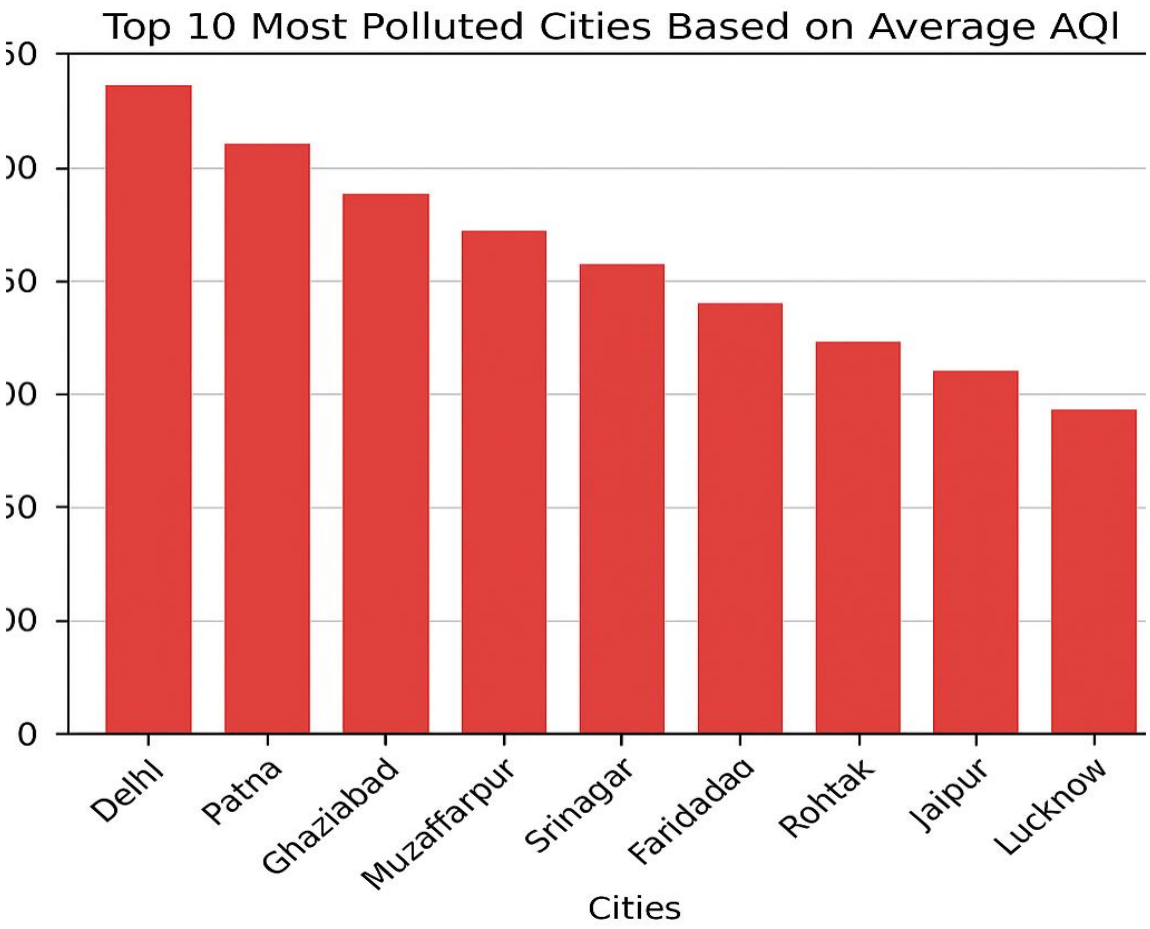
•Feature Importance: Showed PM2.5, PM10, and NO2 as top predictors.

•Residual Plot: Random Forest had randomly scattered residuals, indicating good fit.

•Prediction vs Actual: Graph showed close alignment between predicted and actual AQI values.

•Distribution of Errors: Most errors were centered around zero.

•Model Interpretation: The model generalizes well and can be deployed for AQI forecasting.



11.Tools and Technologies Used

•The project used widely adopted tools and libraries in the data science ecosystem:

•Programming Language: Python 3.x

•Notebook Environment: Jupyter Notebook and Google Colab

°Libraries:

•Pandas and numpy for data handling

•Seaborn and matplotlib for visualization

•Scikit-learn for modeling

•Joblib for model serialization

•Version Control: GitHub for source code management

12.Team Members and Contributions

Name	Responsibility
R. Vishnu Kumar	Data Cleaning, EDA, Model Tuning
Subaitha	Feature Engineering, Documentation
Vishal	Model Building and Evaluation
Asina Banu.	Visualization, Final Report

13.conclusion

•This project aimed to build a predictive model for AQI using pollutant concentrations. After extensive preprocessing, EDA, and modeling, Random Forest Regressor emerged as the most effective model. It achieved high accuracy with an R^2 score of 0.90 and significantly lower error metrics compared to baseline models.

•The project demonstrates the potential of machine learning in environmental monitoring and lays the foundation for real-time AQI prediction systems. Accurate predictions allow citizens to make informed decisions and help policymakers implement effective air quality management strategies

14.Future Scope

•Real-time Data Integration: Connect to real-time pollutant APIs for live AQI predictions.

•Weather Data Integration: Add temperature, humidity, and wind speed to improve model accuracy.

•Deployment: Create a web or mobile dashboard using Flask or Streamlit.

•Advanced Models: Experiment with XGBoost, LSTM (for time-series), or hybrid models.

•Multi-city Analysis: Scale the model for different cities and states in India for broader a dataset