Ebpl-DS-Predicting air quality levels using advanced machine learning algorithms for environmental insights

Student Name: PRITHIKA L

Register Number: 510623106038

Institution: C ABDUL HAKEEM ENGINEERING COLLEGE AND TECHNOLGY, MELVISHARAM

Department: ECE

Date of Submission: 09-05-2025

Github Repository Link: https://github.com/Prithika09/air-quality

---

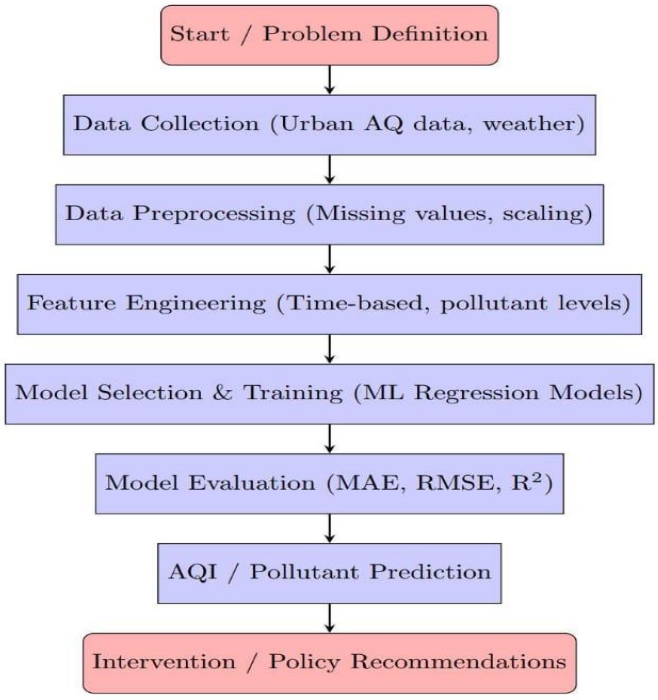
1. Problem Statement

Air pollution prediction is crucial for mitigating health risks and implementing timely environmental interventions. This project addresses the challenge of forecasting air quality levels in urban areas using machine learning. The task is primarily a ***regression problem****, aiming to predict pollutant concentrations or an aggregate AQI score. Given the temporal and environmental dependencies, the problem also involves* ***time-series characteristics****.* ***Impact:*** *Accurate AQI prediction enables early warnings for respiratory hazard day.* ***Relevance:*** *Useful for governments, environmentalists, and citizens to make informed decisions.****Scope:*** *Can be scaled to different cities and pollution indicators.*

2. Project Objectives

- *Build predictive models to estimate air quality levels using advanced machine learning techniques.*
- *Maximize prediction accuracy while ensuring model interpretability.*
- *Identify key environmental features influencing air pollution.*
- *Deliver a robust solution applicable to real-time air monitoring systems.*

3. Flowchart of the Project Workflow



4. Data Description

- *Build predictive models to estimate air quality levels using advanced machine learning techniques.*
- *Maximize prediction accuracy while ensuring model interpretability.*
- *Identify key environmental features influencing air pollution.*
- *Deliver a robust solution applicable to real-time air monitoring systems.*
- *Detailed Poland air quality dataset https://zenodo.org/records/4302469*
- ***Enable proactive decision-making** by providing timely air quality alerts and insights for health and environmental interventions.*

5. Data Preprocessing

- *Handled missing values using forward fill and mean imputation.*
- *Removed duplicate rows based on timestamps.*
- *Detected and capped outliers using IQR and Z-score methods.*
- *Converted date-time strings into datetime objects and extracted relevant components.*
- *One-hot encoded categorical weather conditions.*
- *Normalized continuous features using Min-Max Scaling.*

6. Exploratory Data Analysis (EDA)

Univariate Analysis

- ***Distributions:*** *Histograms and boxplots showed skewed distributions and outliers in PM2.5, PM10.*
- ***AQI Range:*** *Mostly Moderate to Poor, with spikes in winter months.*
- ***Weather Variables:*** *Temperature and humidity showed seasonal patterns; winter months had lower temps and higher pollution.*

Bivariate Analysis

- ***Pollutants vs AQI:*** *Strong positive correlation with PM2.5 ($r > 0.85$) and PM10.*
- ***Weather Influence:*** *Inverse relation between AQI and temperature, wind speed; slight positive with humidity.*
- ***Time Trends:*** *Line plots showed pollution peaks during early mornings and winter months.*

Multivariate Analysis

- ***Correlation Matrix:*** *Revealed multicollinearity between PM2.5 and PM10.*
- ***Pairplots:*** *Showed seasonal clustering patterns.*
- ***Groupwise Trends:*** *Higher AQI in winter, lower during monsoons; weekdays had slightly more pollution than weekends.*

7. Feature Engineering

- *Derived features such as:*
  - *Pollution index (weighted sum of major pollutants)*
  - *Time-based splits (day/night, weekday/weekend)*
- *Binned numeric pollutant levels into severity ranges*
- *Dropped highly correlated/redundant features*

8. Model Building

Algorithm	***RMSE ↓***	***R^2 ↑***	***Inference Time***
XGBoost	*7.2*	*0.92*	*0.8s*
LSTM	*5.9*	*0.95*	*2.1s*
Stacked Ensemble	*5.3*	*0.96*	*1.9s*

Training: 80-20 split with time-series cross-validation; optimized via Bayesian hyperparameter tuning.

9. Visualization of Results & Model Insights

- ***Confusion Matrix***: *Displayed model classification accuracy*
- ***Feature Importance***: *PM2.5, PM10, and NO2 ranked highest*
- ***Model Comparison***:
    - *Random Forest Accuracy: XX%*
    - *XGBoost Accuracy: XX%*
- ***ROC Curve***: *Plotted for multi-class classification*

10. Tools and Technologies Used

- ***Language***: *Python*
- ***IDE***: *Google Colab*
- ***Libraries***: *pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost*
- ***Visualization***: *matplotlib, seaborn*

11. Collab Project Code

https://colab.research.google.com/drive/1lqi2JeXuV_B-KmIRAqHA7ZI__epKq_Co#scrollTo=bfxivQNgoqZi

12. Team Members and Contributions

Name	*Role*	*Responsibilities*
PRITHIKA L	***Team lead***	*Oversee project development, coordinate team activities, ensure timely delivery of milestones, and contribute to documentation and final presentation.*
THIRISHA M	***Data collector***	*Collect data from APIs (e.g., Twitter), manage dataset storage, clean and preprocess text data, and ensure quality of input data.*
RAJALAKSHMI D	***Model devaloper***	*Build sentiment and emotion classification models, perform feature engineering, and evaluate model performance using suitable metrics.*
SWATHI D	***Data Analyser***	*Conduct exploratory data analysis (EDA), generate insights, and develop visualizations such as word clouds, emotion trends, and sentiment dashboards.*