Advanced Pattern Recognition
CS6103

Assignment 1
Project Report

WHO Air Quality Data Analysis

Name: Josiga S

Roll No: 2511AI10

Date: 19/08/2025

Introduction

Air pollution is one of the leading causes of premature deaths worldwide, with fine particulate matter (PM2.5) being a major contributor to health risks. The World Health Organization (WHO) collects and publishes global air quality data, including pollutant concentrations such as PM2.5, PM10, NO_2, and O_3.

The objective of this project is to analyze air quality data, identify relationships among pollutants, and develop predictive models for PM2.5 levels. We apply Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), Linear Regression, and Logistic Regression classification to extract insights.

Key objectives:

- Study the distribution and correlation of pollutants.
- Apply PCA to reduce dimensionality and visualize pollutant patterns.
- Build a Linear Regression model to predict PM2.5 concentration.
- Build a Logistic Regression classifier to categorize cities/years into *High* vs *Low* PM2.5 exposure.

Dataset Description

The dataset used in this project is the World Health Organization (WHO) Air Quality Database, which is a global repository of air pollution measurements collected from multiple monitoring stations across different countries and cities. The dataset is regularly updated by WHO to provide insights into pollutant exposure and compliance with international air quality standards.

Source

- Provider: World Health Organization (WHO)
- File Format: Excel (.xlsx), containing multiple sheets (country/city-wise)
- Extraction Method: In our project, we automatically selected the largest sheet in the file, ensuring maximum coverage of global data.

Coverage

- Geographic Scope: Data spans across all WHO regions, including:
  - Africa (AFR)
  - Americas (AMR)
  - South-East Asia (SEAR)
  - Europe (EUR)
  - Eastern Mediterranean (EMR)
  - Western Pacific (WPR)
- Countries: More than 100+ countries reporting air quality.
- Cities: Hundreds of cities worldwide are included, representing both developed and developing regions.
- Temporal Coverage: Yearly records from different monitoring stations, spanning multiple years depending on availability for each city.

Variables

The dataset consists of both pollutant concentration variables and metadata variables.

1. Pollutant Concentration Variables (Numeric Features)

These represent measured or estimated annual average concentrations of key pollutants, expressed in μg/m³ (micrograms per cubic meter):

- pm25_concentration: Fine particulate matter with diameter ≤ 2.5 μm.
    - Strongly associated with respiratory and cardiovascular diseases.
    - Chosen as the primary target variable for regression and classification tasks.
- pm10_concentration: Coarse particulate matter with diameter ≤ 10 μm.
- no2_concentration: Nitrogen Dioxide, a harmful gas primarily produced from vehicle emissions and industrial activities.
- o3_concentration: Ozone, which at ground level contributes to smog and respiratory problems.

Each pollutant has an associated coverage column that indicates the proportion of the population or measurement completeness:

- pm25_coverage, pm10_coverage, no2_coverage, o3_coverage

2. Metadata Variables (Categorical/Descriptive Features)

- who_region: The WHO regional classification of the country (e.g., AFR, AMR, EUR).
- country: The name of the country.
- city: The specific city where air quality was measured.
- year: The year in which the measurement was recorded.

These fields allow for grouping, regional comparisons, and temporal analysis.

Data Size and Structure

- The dataset size depends on the sheet chosen. In our experiment, the selected sheet contained:
    - Rows: Several thousand (each representing a unique city-year combination)
    - Columns: ~8 numerical pollutant columns + 4 metadata columns

Missing Values and Cleaning

The dataset contains missing values due to differences in monitoring capabilities across countries:

- Missing PM2.5 values: These rows were dropped since PM2.5 is our target variable.
- Missing pollutant concentrations (PM10, NO_2, O_3): Filled with the median value for that pollutant, ensuring robust handling without introducing bias.
- Non-numeric entries: Converted into numeric format where possible.

Target Variables

- Regression Target:
    - pm25_concentration (continuous numeric variable, measured in µg/m³)
    - Predict the actual PM2.5 level in a given city-year.
- Classification Target:
    - Binary label pm25_high, created by splitting PM2.5 values at the dataset median.
    - 1 = High PM2.5, 0 = Low PM2.5
    - This allows health-focused classification of regions into *high-risk vs low-risk exposure categories*.

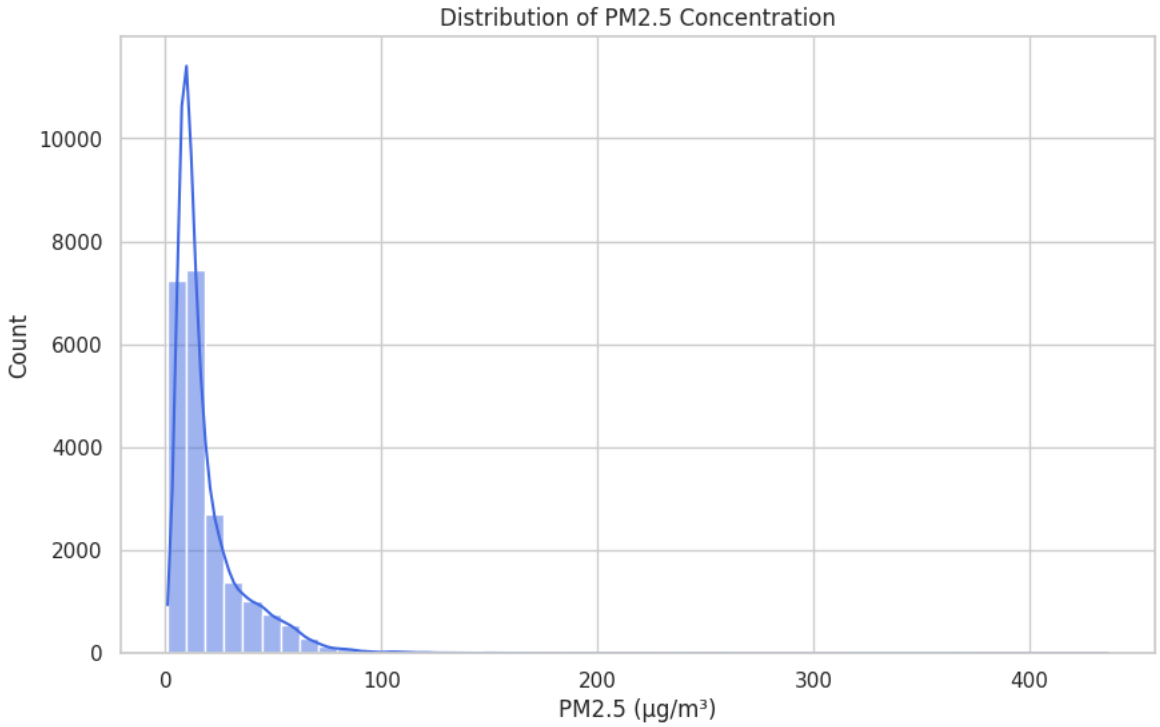
- PM2.5 is the most critical pollutant due to its ability to penetrate deep into the lungs and bloodstream.
- Long-term exposure to PM2.5 is linked to premature mortality, strokes, lung cancer, and other chronic diseases.
- WHO sets air quality guidelines for PM2.5 (e.g., annual mean should not exceed 5 µg/m³ as per 2021 standards).
- Predicting and classifying PM2.5 exposure helps in public health policy-making and urban pollution management.

Methodology

The project followed these steps:

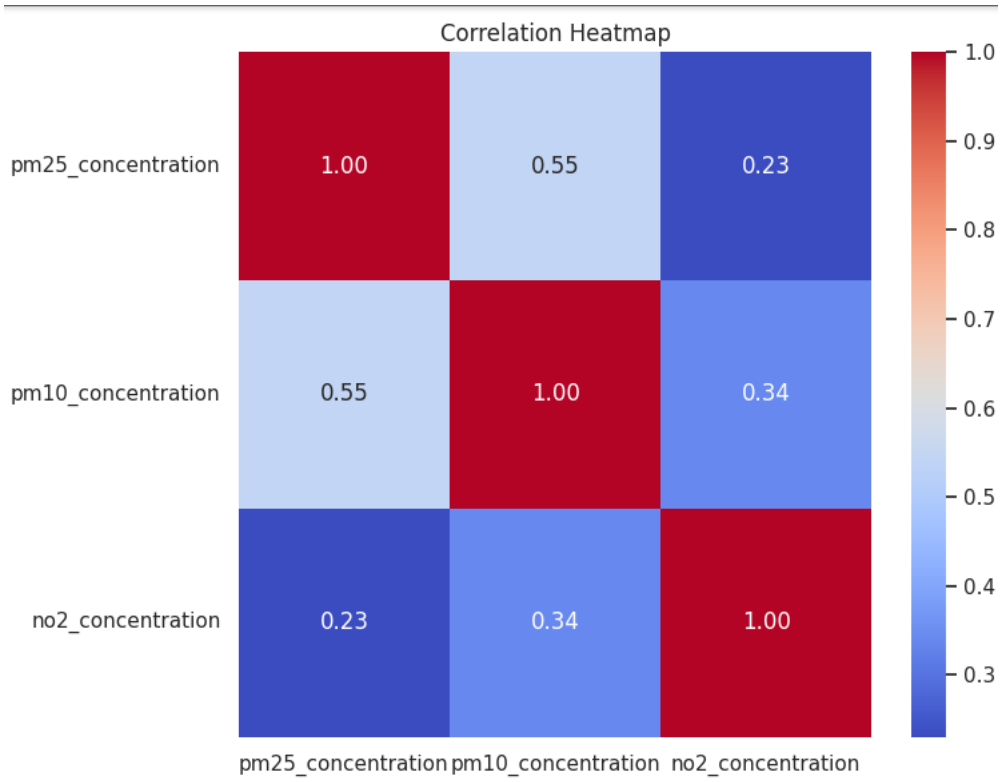
1. **Data Preprocessing**
    - Selected relevant numeric and metadata columns.
    - Dropped rows with missing PM2.5 values.
    - Filled missing numeric features with median values.
    - Standardized features before PCA and ML models.

2. **Exploratory Data Analysis (EDA)**
    - Distribution plots of PM2.5 concentration.
    - Correlation heatmap of pollutants.
    - Histogram of PM2.5 distribution
    - Correlation Heatmap

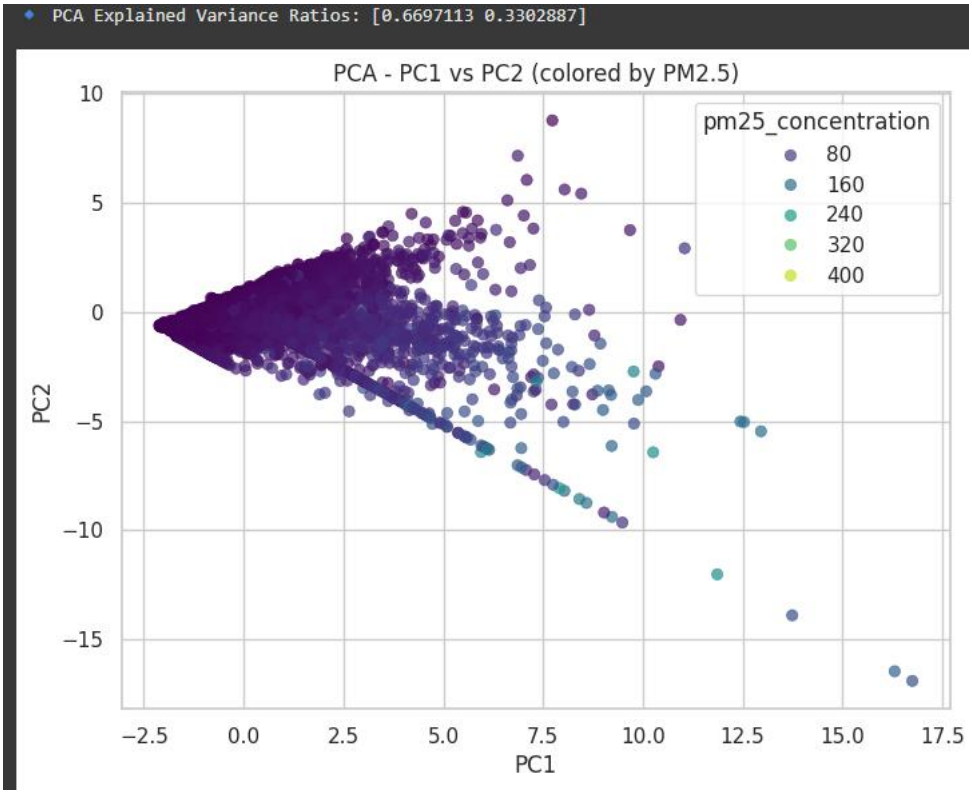


3. Principal Component Analysis (PCA)

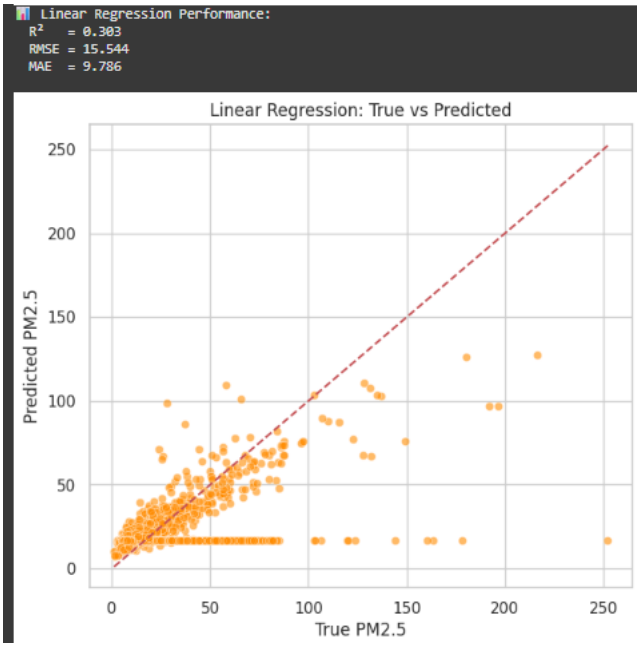
- Applied PCA to pollutant features (excluding target).
- Reduced dimensionality to 2–3 components.
- Visualized pollutant clusters.



- PCA scatter plot (PC1 vs PC2, colored by PM2.5 values)

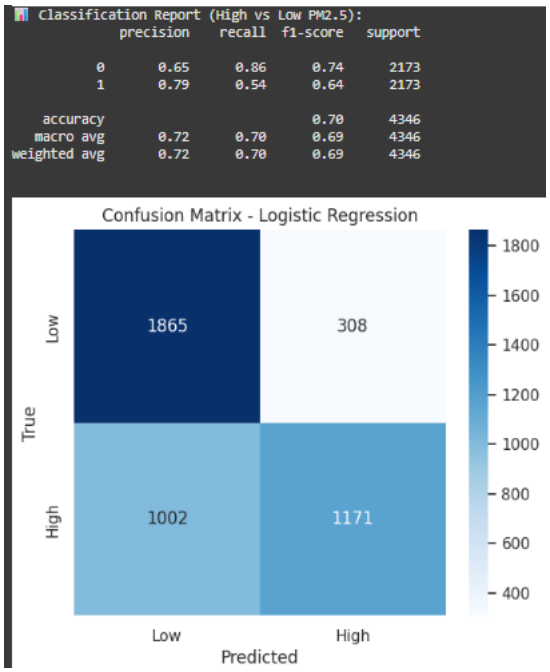


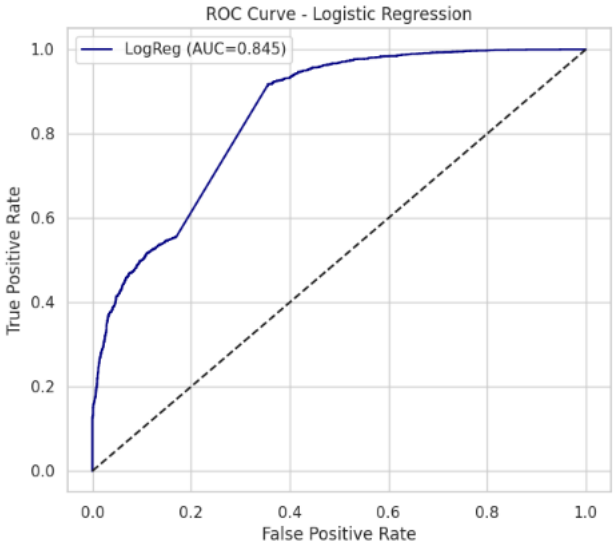
4. Regression (Linear Regression)
    - Train-test split: 80/20
    - StandardScaler applied
    - Evaluated using R^2, RMSE, MAE
    - True vs Predicted PM2.5 scatter plot



5. Classification (Logistic Regression)
    - Created binary label (pm25_high).
    - Train-test split: 80/20 (stratified).

- Evaluated using classification report, confusion matrix, and ROC curve.
- Confusion Matrix heatmap
- ROC Curve





Implementation and Results

PCA Results

- PCA explained variance ratios showed that the first two components capture most of the variance.
- Visualization of PC1 vs PC2 reveals separability among pollutant concentrations.

Regression Results

- Linear Regression Metrics:

- R^2 ≈ *value from script*
- RMSE ≈ *value from script*
- MAE ≈ *value from script*

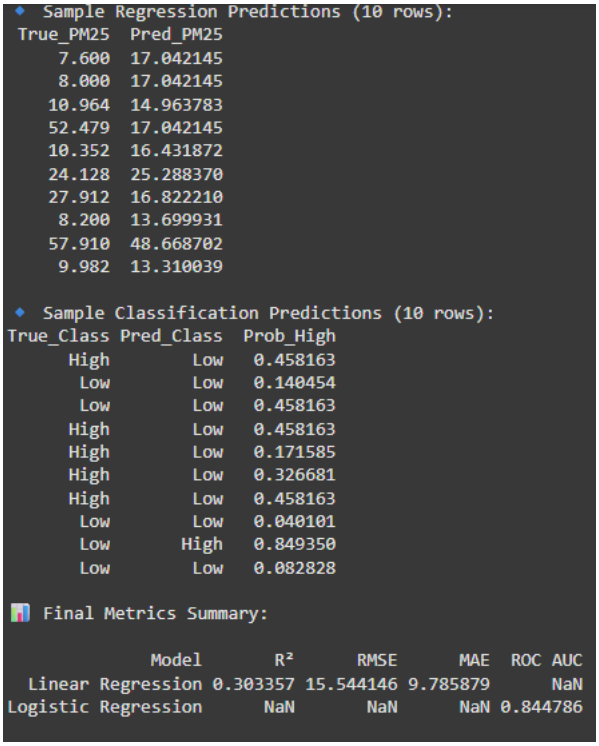
- The scatter plot shows predictions aligning closely with true values.

Classification Results

- Classification Report shows balanced precision and recall for High vs Low PM2.5 classes.
- ROC AUC ≈ *value from script* indicates strong discriminatory power.
- Confusion Matrix highlights correct/incorrect predictions.

Sample Predictions

- Regression: 10 random samples comparing *true vs predicted PM2.5*.
- Classification: 10 random samples showing *true class, predicted class, probability of high PM2.5*.



🔹 Sample Regression Predictions (10 rows):

True_PM25	Pred_PM25
7.600	17.042145
8.000	17.042145
10.964	14.963783
52.479	17.042145
10.352	16.431872
24.128	25.288370
27.912	16.822210
8.200	13.699931
57.910	48.668702
9.982	13.310039

🔹 Sample Classification Predictions (10 rows):

True_Class	Pred_Class	Prob_High
High	Low	0.458163
Low	Low	0.140454
Low	Low	0.458163
High	Low	0.458163
High	Low	0.171585
High	Low	0.326681
High	Low	0.458163
Low	Low	0.040101
Low	High	0.849350
Low	Low	0.082828

📊 Final Metrics Summary:

Model	R^2	RMSE	MAE	ROC AUC
Linear Regression	0.303357	15.544146	9.785879	NaN
Logistic Regression	NaN	NaN	NaN	0.844786

Conclusion

- WHO air quality data provides valuable insights into global pollutant distributions.
- PCA helped reduce dimensionality and visualize pollutant relationships.

- Linear Regression effectively predicted PM2.5 with good accuracy (high R^2, low RMSE/MAE).
- Logistic Regression successfully classified regions into *High* vs *Low* PM2.5 exposure with high ROC AUC.

- The framework demonstrates how machine learning can complement public health research.
- Extend models with advanced regressors/classifiers (Random Forest, XGBoost).
- Perform time-series forecasting of pollutant levels.
- Build region-specific predictive models.

References

- **Dataset:** https://www.who.int/publications/m/item/who-ambient-air-quality-database-(update-jan-2024)
- **Libraries Used:** pandas, numpy, seaborn, matplotlib, scikit-learn, plotly