



Semester: Fall 2025

Deadline: 28-Dec-2025

Semester Project

CSC380 Introduction to Data Science

Total Marks: 25

Weightage: 15%

Problem Statement

Global Air Quality Dataset: Comprehensive Air Quality Measurements from Major Cities Worldwide

Problem Description

Air pollution is a global concern. The Global Air Quality Data dataset (10,000 records) provides an extensive compilation of air quality measurements from various prominent cities worldwide. This dataset includes crucial environmental indicators such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), carbon monoxide (CO), and ozone (O₃), along with meteorological data like temperature, humidity, and wind speed.

The dataset is composed of the following columns:

City: The name of the city where the air quality measurement was taken.

Country: The country in which the city is located.

Date: The date when the measurement was recorded.

PM2.5: The concentration of fine particulate matter with a diameter of less than 2.5 micrometers ($\mu\text{g}/\text{m}^3$).

PM10: The concentration of particulate matter with a diameter of less than 10 micrometers ($\mu\text{g}/\text{m}^3$).

NO₂: The concentration of nitrogen dioxide ($\mu\text{g}/\text{m}^3$).

SO₂: The concentration of sulfur dioxide ($\mu\text{g}/\text{m}^3$).

CO: The concentration of carbon monoxide (mg/m^3).

O₃: The concentration of ozone ($\mu\text{g}/\text{m}^3$).

Temperature: The temperature at the time of measurement ($^\circ\text{C}$).

Humidity: The humidity level at the time of measurement (%).

Wind Speed: The wind speed at the time of measurement (m/s).

Your task:

- Analyze air quality trends/patterns
- Build predictive models for AQI or pollutant levels.
- Understand the impact of pollution on health
- Suggest strategies for environmental improvement

Instructions to Students

1. Work individually. Each student must complete the CCP on their own.
2. Use Python with the following libraries: Pandas, NumPy, Matplotlib/Seaborn, and Scikit-learn.
3. Download the Global Air Quality Dataset from Kaggle (<https://www.kaggle.com/datasets/waqi786/global-air-quality-dataset/data>).
4. Check and document any data limitations, such as missing values, missing features, incorrect readings, or inconsistent measurements.
5. Submit your PDF report online through Google Classroom by 28 December 2025.
6. Bring a printed copy of your report for viva.
7. You will NOT receive marks if you cannot clearly explain your work, results, or code during the viva.

Deliverables

1. Data Preprocessing & Cleaning
 - a. Handle missing values (e.g., using interpolation or imputation).
 - b. Detect and handle outliers (using IQR or Z-score).
 - c. Encode AQI (or main target variable) into categories such as: Good, Moderate, Unhealthy, Hazardous (if you are doing classification).
 - d. Scale numerical features (using Normalization or Standardization, where needed).
 - e. Split the dataset into 80% training and 20% testing.
2. Exploratory Data Analysis (EDA)
 - a. Univariate Analysis
 - b. Bivariate Analysis
 - c. Correlation Analysis
 - d. Comparative Analysis
 - e. Identify cycles (if any)
3. Model Building & Prediction
 - a. Apply at least FIVE (5) machine learning algorithms of your choice.
 - b. Evaluate each model using appropriate metrics.
4. Model Interpretation and Recommendations
 - a. Prepare a technical report based on the findings to describing impact of pollutants on health, and suggest strategies for environmental improvement.

Criteria	Excellent (5)	Good (4)	Satisfactory (3)	Needs Improvement (2)	Poor (1)
Data Preprocessing & Cleaning (A2, A3)	Data is clean, missing values and outliers handled appropriately, limitations documented	Minor issues in cleaning, but handled enough for modeling	Basic cleaning; some gaps remain but model runs	Poor cleaning; missing values or outliers ignored	No meaningful cleaning
EDA & Statistical Analysis (A2, A3)	Clear and insightful visualizations, patterns identified, correlations analyzed deeply	Good analysis and visualizations; some patterns described	Basic descriptive stats + simple plots	Limited EDA, unclear or superficial	No meaningful EDA
Model Implementation (A3, A8)	Multiple models implemented correctly;	Models implemented correctly; one task done well	Only one model; minimal comparison	Model flawed or only trivial modeling	No working model
Model Evaluation & Metrics (A3)	Robust evaluation: all relevant metrics, comparison across models, reasoning about results	Good evaluation; most metrics used correctly	Limited metrics or partial evaluation	Incorrect or insufficient evaluation	No evaluation or wrong metrics
Interpretation & Recommendations (A2)	Strong, actionable insights; connection to real-world/environmental implications; limitation discussion	Good interpretation; some real-world suggestions	Basic interpretation ; weak link to real world	Weak or vague interpretation	No interpretation or irrelevant conclusions

CCP Attributes Covered

Attribute	Alignment / Justification
A2 – Depth of analysis required	The dataset is global and heterogeneous; relationships between pollutants, weather, seasonal or spatial factors and Air Quality Index (AQI) are complex and non-obvious.
A3 – Depth of knowledge required	Requires use of EDA, statistical reasoning, ML modeling, handling missing/irregular data, feature engineering.
A8 – Interdependence	Subtasks include data cleaning & preprocessing, EDA, modeling (regression/classification), evaluation, interpretation, and recommendation. Each phase depends on earlier steps.