📄 **Global Pollution Analysis – Final Project Report**

---

1. Objective

The goal of this project is to analyze global pollution data to understand patterns and predict energy recovery from pollution levels. This includes building regression models to estimate energy recovery and classify countries based on pollution severity.

---

2. Data Overview

Dataset: Global_Pollution_Analysis.csv
It contains records with pollution indices, energy metrics, industrial waste, and country/year identifiers.

---

3. Data Preprocessing

- Encoded categorical variables like 'Country' using LabelEncoder.
- Normalized continuous variables like 'Air_Pollution_Index', 'CO2_Emissions', and 'Industrial_Waste' using StandardScaler.
- Created a new column 'Pollution_Category' with 3 classes:
    - Low (0), Medium (1), High (2).

---

4. Exploratory Data Analysis (EDA)

- Descriptive statistics showed significant variation in pollution levels.
- Correlation heatmap showed strong positive relationships between Air Pollution Index, CO2 Emissions, and Energy Recovery.
- Bar plots and line plots revealed rising trends in industrial waste and pollution.

---

5. Feature Engineering

- Extracted year-based trends to visualize pollution changes over time.
- Used 'Energy_Consumption_Per_Capita' to get more granular insights.

6. Linear Regression Model – Predicting Energy Recovery

Objective: Estimate the amount of energy that can be recovered based on pollution levels and industrial waste.
Features Used:

- **Air_Pollution_Index**
- **CO2_Emissions**
- **Industrial_Waste**

Target:

- **Energy_Recovered (in GWh)**

Model Results:

- **R^2 Score: -0.025**
- **Mean Squared Error (MSE): 1.15**
- **Mean Absolute Error (MAE): 0.97**

Interpretation:
The R^2 score is negative, indicating that the model performs worse than a horizontal mean line. The prediction of energy recovery based on these features is currently not reliable and may require either additional features, better preprocessing, or model tuning.

7. Logistic Regression Model – Classifying Pollution Severity

Objective: Classify countries into pollution severity categories: Low, Medium, or High.
Features Used:

- **Air_Pollution_Index**
- **CO2_Emissions**

Target:

- **Pollution_Category (0 = Low, 1 = Medium, 2 = High)**

Model Results:

- **Accuracy: 1.00**

- **Precision: 1.00**
- **Recall: 1.00**
- **F1 Score: 1.00**

Interpretation:
The classification model is performing perfectly on the test data, achieving 100% in all metrics. This could mean either the data is very well-separated, or the model may be overfitting if the dataset is small or imbalanced — further validation with new/test data is recommended.

---

8. Insights and Recommendations

- High pollution levels often lead to greater energy recovery potential.
- Air Pollution Index is a reliable indicator of severity.
- Recommend adopting pollution-to-energy technologies in highly polluted regions.
- Encourage governments to support renewable energy solutions.

---

9. Final Deliverables

- Python Jupyter Notebook (.ipynb) with full code, visualizations, and analysis.
- This PDF report summarizing all steps and insights.
- Confusion matrices and graphs shown in the notebook.