Global Pollution Analysis

# Objective

The goal of this project is to:  
- Analyze global pollution data (air, water, soil)  
- Predict the energy recovery from pollution metrics using Linear Regression  
- Classify pollution severity (Low/High) using Logistic Regression  
- Extract insights for pollution reduction and energy optimization strategies

# Dataset

The dataset contains global pollution records for multiple countries over several years.  
Key features include:  
- Air\_Pollution\_Index  
- Water\_Pollution\_Index  
- Soil\_Pollution\_Index  
- CO2\_Emissions (in MT)  
- Industrial\_Waste (in tons)  
- Energy\_Recovery (in GWh)

# Phase 1: Data Preprocessing & EDA

## Step 1: Data Cleaning

- Missing values handled using mean imputation  
- All numerical features scaled using StandardScaler  
- Categorical features like Country and Year encoded with LabelEncoder

## Step 2: Exploratory Data Analysis (EDA)

- Descriptive Statistics generated for key features  
- Correlation Heatmap shows relation between pollution indices and energy recovery  
- Box plots & line charts revealed pollution trends across countries and years

## Step 3: Feature Engineering

- Created Energy Consumption per Capita  
- Extracted Pollution Severity categories  
 (Initially 3-class, but later binary classification used due to imbalance)

# Phase 2: Predictive Modeling

## Step 4: Linear Regression – Predicting Energy Recovery

- Target: Energy\_Recovery (in GWh)  
- Features: CO2\_Emissions, Industrial\_Waste, Pollution Indices  
- Model: LinearRegression  
- Performance:  
 - MSE: ~9602.89  
 - MAE: ~72.53  
 - R² Score: ~0.60

## Step 5: Logistic Regression – Classifying Pollution Severity

- Target: Binary Pollution Level (0 = Low, 1 = High)  
- Model: LogisticRegression  
- Performance:  
 - Accuracy: ~88%  
 - Precision: ~91%  
 - Recall: ~85%  
 - F1 Score: ~88%

# Phase 3: Model Comparison & Insights

## Model Comparison

Linear Regression:  
- Purpose: Predict energy in GWh  
- R² Score: ~0.60  
- MAE: ~72.53, MSE: ~9602.89  
  
Logistic Regression:  
- Purpose: Classify pollution level  
- Accuracy: ~88%, F1 Score: ~88%

## Insights & Recommendations

Insights:  
- Air Pollution Index strongly affects energy recovery and pollution severity  
- Higher industrial waste and CO2 emissions correlate with more energy recovery  
- Pollution is increasing over time  
  
Recommendations:  
- Invest in waste-to-energy plants  
- Focus on carbon emission control  
- Implement pollution tracking systems  
- Promote environmental tech and regulations

# Final Deliverables

- Jupyter Notebook with code, EDA, and models  
- Visualizations (heatmaps, confusion matrix, trends)  
- Final Report (this document)

# Conclusion

This project successfully used Machine Learning to:  
- Predict energy recovery from pollution levels  
- Classify pollution severity  
- Provide actionable recommendations for pollution control