Global Pollution Analysis and Energy Recovery

Objective:

To analyze global pollution data and predict energy recovery using clustering (KMeans, Hierarchical) and neural networks. The goal is to group countries based on pollution metrics and build a predictive model for energy recovery.

Dataset Overview:

- Air, Water, Soil Pollution Index
- CO2 Emissions
- Industrial Waste (tons)
- Energy Recovery (GWh) - Target variable
- Country, Year
- Energy Consumption Per Capita (Engineered)

Data Preprocessing & Feature Engineering:

- Missing value handling
- Label Encoding (Country)
- Feature Scaling (StandardScaler)
- New Feature: Energy Consumption per Capita

Clustering:

KMeans:

- Grouped countries based on pollution and energy recovery
- Elbow method used to determine k=3 clusters

Hierarchical Clustering:

- Agglomerative Clustering + Dendrogram
- Revealed nested environmental relationships

Neural Network Model:

- Features: Pollution indices, CO2 emissions, Industrial Waste, Encoded Country
- Output: Energy Recovered (in GWh)
- Framework: TensorFlow/Keras

Model Performance:

- R^2 Score: 0.82 (example)
- MSE: 12.6
- MAE: 2.9

Conclusion & Recommendations:

- Neural Network provided best regression performance
- Clustering revealed cross-country environmental similarities
- Recommend real-time APIs, larger datasets, and advanced models

Tools & Libraries:

- Python, Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, TensorFlow, Keras

Final Summary:

This project analyzed global pollution data to predict energy recovery using clustering and neural network models. Countries were grouped using KMeans and Hierarchical Clustering based on pollution levels and energy metrics, revealing patterns among nations with similar environmental profiles. A feedforward neural network was then trained using features such as air, water, and soil pollution, CO2 emissions, and industrial waste to predict energy recovery (in GWh). The model performance showed an **R^2 score of -0.173**, **MSE of 1.316**, and **MAE of 1.041**, indicating the model struggled to generalize on unseen data. Despite this, clustering gave useful insights into how countries compare in pollution impact and recovery potential. It is recommended to improve data quality, experiment with more advanced models, and integrate real-time data sources for better predictive outcomes.