Pollution Severity Classification using Machine Learning

🟩 1. Objective

The main objective of this machine learning project is to classify countries into pollution severity levels — specifically, categories of High and Low — based on environmental, energy, and industrial parameters. The task involves analyzing a real-world dataset that includes various indicators such as air, water, and soil pollution indices, CO_2 emissions, energy consumption per capita, and industrial waste generation.
The ultimate goal is to help policymakers and researchers automatically identify pollution severity in different countries and propose mitigation strategies based on predicted patterns.

🟩 2. Dataset Overview

The dataset used in this project is titled Global_Pollution_Analysis.csv, which contains multi-year records of country-level environmental statistics. Key attributes in the dataset include:

- Air_Pollution_Index, Water_Pollution_Index, Soil_Pollution_Index
- CO2_Emissions (in MT), Industrial_Waste (in tons)
- Energy_Consumption_Per_Capita (in MWh), Renewable_Energy (%)
- Country, Year, and Population

A new target variable, Pollution_Severity, was derived by aggregating the three pollution indices (air, water, and soil) into a Total_Pollution_Index, followed by threshold-based classification into two categories: Low and High pollution severity.

🟩 3. Data Preprocessing

To ensure robust model performance, the following preprocessing steps were applied:

- Missing value treatment: Numerical columns with nulls were imputed using the mean strategy.
- Categorical encoding: The Country and Year columns were label-encoded.
- Scaling: All numerical features were normalized to the [0, 1] range using MinMaxScaler, as required by the Naive Bayes classifier.
- Target encoding: The Pollution_Severity column was label-encoded into binary format: 0 = High, 1 = Low.

Additionally, features like Energy_Consumption_Per_Capita, Industrial_Waste, and CO2_Emissions were selected as key predictors for severity classification.

🟩 4. Feature Engineering

Feature engineering included:

- Total Pollution Index: A derived feature created by summing air, water, and soil pollution scores.
- Binary Classification Labeling: Countries were labeled as 'High' or 'Low' based on threshold ranges in total pollution index.
- Normalization: Features such as emissions and energy per capita were normalized to ensure all variables contributed equally to model learning.

These steps ensured more meaningful learning patterns were extracted by the models.

🟩 5. Model Evaluation and Results

🔷 5.1 Multinomial Naive Bayes Classifier

- Accuracy: 77.5%
- Confusion Matrix:

[[28 1]

[8 3]]

- Classification Report:

	precision	recall	f1-score	support
High	0.78	0.97	0.86	29
Low	0.75	0.27	0.40	11

Analysis:
The Naive Bayes classifier achieved 77.5% accuracy. It performed well on identifying high pollution severity with a high recall of 97%, indicating it could correctly identify almost all high-pollution countries. However, it struggled with the Low class, which had a low recall of only 27%, likely due to class imbalance. The weighted F1-score of 0.73 indicates reasonably balanced overall performance. This suggests Naive Bayes is sensitive to class frequency and may benefit from class balancing.

🔷 5.2 K-Nearest Neighbors (KNN)

- Best k: 5
- Accuracy: 80%
- Confusion Matrix:

[[21 3]

[5 11]]

- Classification Report

	precision	recall	f1-score	support
High	0.81	0.88	0.84	24
Low	0.79	0.69	0.73	16

Analysis:
KNN achieved 80% accuracy, outperforming Naive Bayes in both precision and recall. It was more balanced, correctly identifying both High and Low severity classes. The best k value was determined via cross-validation and was found to be 5. KNN's strength lies in

its ability to adapt to local structure in the data. However, its performance is sensitive to feature scaling and outliers, making proper normalization essential.

🔷 5.3 Decision Tree Classifier

- Best Parameters: max_depth = 5, min_samples_split = 10
- Accuracy: 85%
- Confusion Matrix:

[[20 4]

 [2 14]]

- Classification Report:

	precision	recall	f1-score	support
High	0.91	0.83	0.87	24
Low	0.78	0.88	0.82	16

Analysis:
The Decision Tree classifier was the best performing model with an accuracy of 85%. It showed strong performance on both classes, with high precision (0.91 for High, 0.78 for Low) and good recall (0.83 and 0.88 respectively). The model was tuned using grid search, and the optimal depth was found to be 5. Decision Trees are easy to interpret and perform well even without feature scaling, which gives them an edge in real-world deployments.

🟩 6. Final Summary

This project evaluated the performance of three machine learning algorithms — Multinomial Naive Bayes, K-Nearest Neighbors, and Decision Tree — in classifying pollution severity of countries using environmental features.
The Decision Tree classifier outperformed others, achieving 85% accuracy, followed by KNN with 80%, and Naive Bayes with 77.5%.

- Naive Bayes, while efficient and probabilistically sound, struggled with the minority class (Low).
- KNN provided a better balance but required fine-tuning for k.
- Decision Tree delivered the most stable and accurate results across both classes and proved to be the most suitable for this binary classification problem.

Recommendations: For future improvements, data augmentation or resampling techniques like SMOTE can be applied to address class imbalance. Additionally, ensemble methods like Random Forest or Gradient Boosted Trees could be explored for even higher accuracy