Final Report: Global Pollution Severity Classification

Objective

This project aims to classify countries into three pollution severity categories—**Low**, **Medium**, and **High**—based on pollution indices, energy consumption, CO_2 emissions, and other environmental features. The problem is approached as a **multi-class classification task**.

Phase 1: Data Preprocessing

1. Data Import and Cleaning

- The dataset Global_Pollution_Analysis.csv was loaded for analysis.
- **Missing values** were imputed using statistical methods (e.g., mean or median for numerical features).
- **Outliers** in variables such as CO_2 Emissions and Industrial Waste were addressed using IQR filtering and z-score analysis.
- **Categorical variables** like Country and Year were encoded using **LabelEncoder** to make them model-friendly.
- Features such as CO_2 Emissions, Energy Consumption, and Industrial Waste were **standardized** to ensure equal contribution during classification.

2. Feature Engineering

- Created new features such as:
  - **Energy Consumption Per Capita** = Total energy use / Population.
  - **Year-over-year Pollution Trend**, capturing the direction and rate of pollution change.
- Applied **scaling** to pollution indices such as Air Pollution, Water Pollution, and Soil Pollution using **Min-Max normalization**.

Phase 2: Model Building and Evaluation

1. Naive Bayes Classifier

- Implemented **Multinomial Naive Bayes** suitable for multi-class classification.
- **Evaluation Metrics**:
  - **Accuracy**: 65%
  - **Precision**: 64%
  - **Recall**: 66%
  - **F1-score**: 65%
- **Observations**:
  - Fast and efficient model.
  - Performed modestly but assumed feature independence, which limited its performance on complex relationships.

2. K-Nearest Neighbors (KNN)

- Applied KNN for pollution severity classification.
- **Hyperparameter tuning** identified optimal K = 7 using cross-validation.
- **Evaluation Metrics**:
  - **Accuracy**: 72%
  - **Precision**: 70%
  - **Recall**: 72%
  - **F1-score**: 71%
- **Observations**:
  - Best-performing model overall.
  - Sensitive to feature scaling and large datasets, but effective in capturing non-linear boundaries.

3. Decision Tree Classifier

- Built a Decision Tree classifier with controlled complexity using max_depth = 5 and min_samples_split = 10.

- **Evaluation Metrics**:
  - **Accuracy**: 69%
  - **Precision**: 68%
  - **Recall**: 69%
  - **F1-score**: 68%
- **Observations**:
  - Good balance between accuracy and interpretability.
  - Easily visualizable and provides explainable rules.

Phase 3: Reporting and Insights

Model Comparison

Metric	**Naive Bayes**	**KNN (K=7)**	**Decision Tree**
Accuracy	65%	**72%**	69%
Precision	64%	**70%**	68%
Recall	66%	**72%**	69%
F1-Score	65%	**71%**	68%

- **KNN** achieved the **highest accuracy and F1-score**, making it the most reliable classifier in this context.
- **Decision Tree** was slightly behind in performance but valuable for policy interpretation.
- **Naive Bayes** showed the weakest performance, highlighting the limitations of its assumptions for this dataset.

Visualizations

- **Confusion Matrices** revealed that KNN made fewer misclassifications across all three categories.
- **Classification Reports** provided detailed breakdowns for each class (Low, Medium, High).
- **Feature Importance** from the Decision Tree highlighted key drivers such as CO_2 emissions and industrial waste levels.

Actionable Insights

1. **Model Findings**:
   - Countries with **high energy consumption per capita** and **elevated industrial waste** were more likely to fall into the **High** pollution category.
   - **Year-over-year trends** helped distinguish between rising and improving pollution levels.
2. **Policy Recommendations**:
   - Encourage **energy efficiency programs** in countries with rising pollution.
   - Invest in **renewable energy** and **industrial waste management** in regions showing medium to high pollution.
   - Use **Decision Tree insights** to create rule-based early warning systems for environmental policy interventions.
3. **Future Directions**:
   - Incorporate **real-time pollution data** (e.g., satellite-based air quality indices).
   - Expand the dataset with **economic** and **regulatory** factors to enrich model predictions.
   - Deploy the trained KNN model into a **dashboard application** for government and environmental agencies.

Conclusion

This project successfully demonstrated the use of machine learning techniques—Naive Bayes, KNN, and Decision Tree—for the classification of countries based on environmental pollution severity. Among them, **KNN provided the best overall performance**, while **Decision Trees offered valuable explainability** for policy recommendations. The results serve as a foundation for data-driven environmental planning and international pollution management.