Global Pollution Analysis Report

# 1. Introduction

This project classifies regions as High or Low Pollution Zones using machine learning. By analyzing CO₂, SO₂, and NO₂ levels, it applies Naive Bayes, KNN, and Decision Tree models to identify the most accurate method for environmental monitoring and pollution prediction.

# 2. Dataset Description

 **Dataset Size:** Medium

 **Features:** Pollution-related metrics like CO₂, SO₂, NO₂, Energy Consumption, Pollution Index, etc.

 **Target Variable:** Binary label representing Pollution Zone (0 = Low, 1 = High)

 **Class Balance:** Mild imbalance observed in target classes

 **Preprocessing:**

* Removed irrelevant columns (e.g., name, email, location)
* Checked and handled missing values
* Encoded categorical variables using label encoding
* Standardized numerical features (important for kNN)
* Dataset split into training (80%) and testing (20%)

# 3. Models & Parameters

|  |  |
| --- | --- |
| Model | Notes on Tuning / Setup |
| Naïve bayes | Used Multimonimal() for handling continuous features |
| Decision tree | Used default criterion (likely Gini);max depth set |
| K NN | Used k=3, normalized input features using StandardScaler() |

# 4. Evaluation Metrics

Metrics used for evaluation:  
 **Accuracy** – Overall correctness of predictions

 **Precision** – How many predicted “high” zones were truly high

 **Recall** – How many actual “high” zones were correctly predicted

 **F1 Score** – Harmonic mean of precision and recall

 **Confusion Matrix** – TP/FP/FN/TN counts for understanding error distribution

# 5. Results

## 5.1 Performance Comparison

| **Model** | **Accuracy** | **Precision** | **F1 Score** | **Confusion Matrix** |
| --- | --- | --- | --- | --- |
| **Naive Bayes** | 0.3 | 0.32 | 0.34 | [[ 5 2 3], [ 6 4 2],[ 10 5 3]] |
| **kNN (k=3)** | 0.3 | 0.319 | 0.28 | [[ 4 3 3], [ 6 3 3],[ 12 2 4]] |
| **Decision Tree** | 1.0 | 1.0 | 1.0 | [[16 0 0], [0 8 0],[ 0 0 8]] |

# 6. Analysis

 **Naive Bayes** achieved the lowest precision and F1 Score, showing non balanced performance.

 **kNN** achieved the lowest precision and F1 Score, showing non balanced performance.

 **Decision Tree** was able to fully capture class boundaries in the dataset..

# 7. Conclusion & Future Work

Best Model: With 100% accuracy and perfect evaluation metrics, the Decision Tree model proved most effective for classifying pollution zones in this dataset.  
Future Work:  
 Apply cross-validation to ensure robustness and avoid overfitting

Use ensemble models like Random Forest and XGBoost

Add temporal and regional pollution trends for real-time modeling