Global Pollution Analysis and Energy Recovery

Objective

The goal of this project is to analyze global pollution data to understand the relationship between pollution levels (air, water, soil) and energy recovery across different countries. By leveraging the **Apriori algorithm**, we aim to uncover hidden associations between pollution factors and energy consumption that can guide strategies for pollution control and energy optimization. `This dataset will be used for` **`Apriori Algorithm`** `and` **`evaluation/validation`** `techniques.`

Phase 1 - Data Preprocessing and Feature Engineering

- **Data Import and Cleaning**
    - Load the dataset (**`Global_Pollution_Analysis.csv`**).
    - **Handle Missing Data** Identify missing or inconsistent values and apply imputation techniques (e.g., mean imputation or removal of rows/columns).
    - **Normalize Pollution Indices** Scale the features related to pollution (air, water, and soil pollution) for consistency.
    - **Encode Categorical Features** Apply label encoding to categorical columns such as `country` and `year` to convert them into numerical values.
- **Feature Engineering**
    - **Energy Consumption per Capita** Derive a new feature for energy consumption per capita to better understand energy efficiency across countries.
    - **Pollution Trends** Analyze the trends in pollution levels over the years to identify potential correlations with energy recovery.
    - **Categorization of Pollution Severity** Create categorical features representing pollution severity (e.g., Low, Medium, High) based on specific thresholds for air and water pollution indices.

Phase 2 - Apriori Algorithm for Market Basket Analysis

1. **Introduction to Apriori Algorithm**
    - **Objective** Apply the Apriori Algorithm to find association rules between pollution levels, energy consumption, and recovery metrics.
    - **Implementation** Use the **Apriori Algorithm** to extract frequent itemsets (e.g., countries with similar pollution patterns and energy recovery rates).
2. **Mining Association Rules**
    - Apply the **Apriori Algorithm** to identify associations between high pollution levels and the type of energy recovery.
    - Set appropriate thresholds for minimum support and confidence.
3. **Evaluation**
    - Visualize frequent itemsets and association rules using graphs.
    - Interpret the rules to understand how pollution levels correlate with energy recovery and make strategic recommendations.

Phase 3 - Model Evaluation and Validation

1. **Model Validation**
    - **Objective:** Make sure the Apriori algorithm's rules are valid and provide meaningful insights.

    - **Implementation:** Split the data into training and test sets, use cross-validation, and evaluate the accuracy of rules mined through Apriori.
2. **Evaluation Metrics**
    - Measure the effectiveness of the Apriori algorithm by assessing the **lift** and **confidence** of the rules.
    - Analyze the **support** for different itemsets to make sure that associations are statistically significant.

Phase 4 - Reporting and Insights

1. **Model Comparison**
    - Compare the performance of CNN models for delivery prediction and Apriori for association rule mining.
    - Discuss the effectiveness of using image-based features for CNN and rule-based learning in Apriori.
2. **Actionable Insights**
    - Based on the CNN and Apriori model outputs, provide recommendations on improving delivery time predictions and optimizing pollution control strategies.
3. **Final Deliverables**
    - **Jupyter Notebook (.ipynb)**
        - Full code for CNN, Apriori, and model validation, with detailed explanations for each step.
    - **Data Visualizations**
        - Visualizations like confusion matrices, ROC curves, and association rule graphs.
    - **Final Report**
        - A comprehensive report summarizing methodologies, evaluations, insights, and actionable recommendations.

Final Deliverables

1. **Jupyter Notebook (.ipynb)**
    - Full code for CNN implementation, Apriori algorithm application, and evaluation/validation methods.
2. **Data Visualizations**
    - Visualizations such as confusion matrix for CNN performance, ROC curve for model evaluation, and association rule graphs for Apriori results.
3. **Final Report**
    - A comprehensive report that covers the methodology, model performance, and key findings from CNN, Apriori, and model validation techniques.
