**Market Basket Analysis using Apriori Algorithm**

**1. Introduction**

The Apriori algorithm is a popular association rule mining technique used to uncover hidden relationships between variables in large datasets. In this project, we applied Apriori to analyze environmental and energy-related factors such as pollution levels, waste production, energy recovery, and renewable energy.

The objective is to:

* Identify **frequent itemsets** (patterns of co-occurring factors).
* Generate **association rules** to understand how one condition (antecedent) is related to another (consequent).
* Validate the rules using training and test splits, ensuring that the associations are robust.

**2. Methodology**

1. **Data Preparation**
   * Continuous variables were categorized into levels (Low, Medium, High).
   * This enabled Apriori to treat them as categorical items.
2. **Apriori Algorithm**
   * Minimum support threshold: selected to capture frequently co-occurring items.
   * Minimum confidence threshold: ensured that only strong predictive associations were considered.
   * Metrics used: *Support, Confidence, Lift*.
3. **Validation**
   * Performed **cross-validation** (training vs test support and confidence).
   * Checked rule stability across different folds.

**3. Results (Top Association Rules)**

| **Antecedent** | **Consequent** | **Support (Train)** | **Confidence (Train)** | **Confidence (Test)** | **Lift** |
| --- | --- | --- | --- | --- | --- |
| (GDP Per Capita Low) | (Renewable Energy High) | 0.163 | 0.500 | 0.333 | 1.455 |
| (Industrial Waste High) | (Water Pollution High) | 0.175 | 0.528 | 0.143 | 1.509 |
| (Industrial Waste Medium) | (CO2 Emissions Low) | 0.175 | 0.519 | 0.167 | 1.455 |
| (Plastic Waste Low) | (Soil Pollution Low) | 0.172 | 0.509 | 0.345 | 1.524 |
| (Population Low) | (Water Pollution Low) | 0.177 | 0.524 | 0.451 | 1.530 |
| (Soil Pollution Low) | (Plastic Waste Low) | 0.175 | 0.538 | 0.267 | 1.566 |
| (Water Pollution Low) | (Population Low) | 0.177 | 0.517 | 0.477 | 1.530 |

**4. Interpretation**

* **Economic vs Renewable Energy**
  + Countries with **low GDP per capita** are often associated with **higher renewable energy usage** (Lift = 1.45).
  + Possible reason: low-GDP countries may depend more on hydro, solar, or biomass due to limited fossil fuel infrastructure.
* **Industrial Waste & Pollution**
  + High industrial waste correlates strongly with **water pollution** (Lift = 1.51).
  + Medium waste links to **low CO2 emissions**, suggesting industries in such cases may dispose of waste improperly instead of energy-efficient incineration.
* **Plastic & Soil Pollution**
  + Low plastic waste aligns with **low soil pollution** (Lift = 1.52).
  + Indicates effective waste management systems where controlling plastic waste also reduces soil contamination.
* **Population & Water Pollution**
  + Lower population correlates with **lower water pollution** (Lift = 1.53).
  + Suggests population density plays a direct role in water quality.

**5. Actionable Insights**

1. **Policy Recommendations**
   * **Target Industrial Waste Reduction**: Since industrial waste is directly linked to water pollution, stricter waste treatment regulations could substantially improve water quality.
   * **Encourage Plastic Waste Control**: Reducing plastic use and improving recycling can help decrease soil contamination.
2. **Energy Strategy**
   * Countries with lower GDP but high renewable adoption should be used as **case studies** for cost-effective renewable integration.
   * Wealthier countries could replicate these low-cost models in rural or underserved regions.
3. **Population Management**
   * Urban planning should include water resource protection, as high population densities contribute to higher pollution levels.
4. **Cross-Metric Correlation**
   * Interventions should not target one metric in isolation. For example, reducing industrial waste also benefits **water quality and CO2 management simultaneously**.

**6. Conclusion**

In this Apriori-based analysis of global environmental and energy data, we discovered several interpretable and actionable associations:-  
Many top rules had **train confidence around 0.50** and **lift between ~1.45–1.57**, indicating modest but meaningful co-occurrence above chance. Notably, **Population\_Low → Water\_Pollution\_Low** (**train confidence 0.524, test confidence 0.451, lift 1.53**) and **Plastic\_Waste\_Low → Soil\_Pollution\_Low** (**train confidence 0.509, test 0.345, lift 1.52**) were relatively stable across cross-validation folds and therefore reliable signals; by contrast, rules such as **Industrial\_Waste\_High → Water\_Pollution\_High** (**train confidence 0.528 → test 0.143**) showed large drops on held-out data and appear less generalizable. From these patterns, we draw practical recommendations: prioritise **industrial waste treatment** to reduce water pollution, strengthen **plastic-waste controls and recycling** to lower soil contamination, and study **low-GDP countries that show high renewable uptake** as case studies for scalable, low-cost clean-energy policies; also, integrate population-aware urban planning to protect water resources.

Overall, the Apriori rules provide clear, human-readable associations that can guide targeted policy and operational interventions, but they should be used together with cross-validated predictive models and domain expertise to confirm causality and design effective implementations.