Assignment 8

Global Pollution Analysis and Energy Recovery

Project Report

---

1. Objective

To explore global pollution data and uncover patterns related to energy recovery using association rule mining (Apriori algorithm).

---

2. Methodology

2.1 Data Preprocessing

- **Missing Values:** Rows with nulls were dropped for clean model input.
- **Normalization:** Scaled pollution indices (Air, Water, Soil) to a [0,1] range using `MinMaxScaler`.
- **Label Encoding:** Converted categorical variables like *Country* and *Year* to numerical codes.
- **Categorical Derivation:**
  - Derived pollution severity levels (Low, Medium, High) using threshold bins on normalized indices.
  - Created a binary feature `High_Energy_Recovery` using the median of `Energy_Recovered (in GWh)`.

2.2 Association Rule Mining (Apriori Algorithm)

- **Approach:**
  - One-hot encoded categorical pollution levels and appended `High_Energy_Recovery` as a binary target.

- Applied the Apriori algorithm to identify frequent itemsets with a support threshold of 0.1.
- Extracted rules using confidence ≥ 0.5 and lift > 1 for interpretability and significance.

- **Output:**
  - Generated strong association rules indicating conditions under which high energy recovery is likely.
  - Constructed bar plots and a network graph for rule visualization.

---

3. Model Evaluation

Association Rules Evaluation

- **Top Rules:**
  - Found rules such as:
    - `{'High'} in Air Pollution` → `High Energy Recovery`
    - `{'High'} in Soil Pollution` ∧ `{'Medium'} in Water Pollution` → `High Energy Recovery`
- **Lift Values:** All top rules had lift > 1, confirming positive association strength.
- **Network Visualization:** Provided a visual representation of the interdependence of variables.

---

4. Key Findings

1. **Pollution Severity Matters:**
   - High Air or Soil Pollution levels are frequently associated with High Energy Recovery initiatives.
2. **Rule Strength:**
   - Top rules demonstrated strong support and lift, indicating meaningful relationships.

---

5. Recommendations

5.1 Policy & Environmental Strategy

- Target countries or regions showing consistent `High` pollution indicators with tailored energy recovery investments.
- Use identified association rules to guide interventions with higher recovery potential.

5.2 Model Improvements

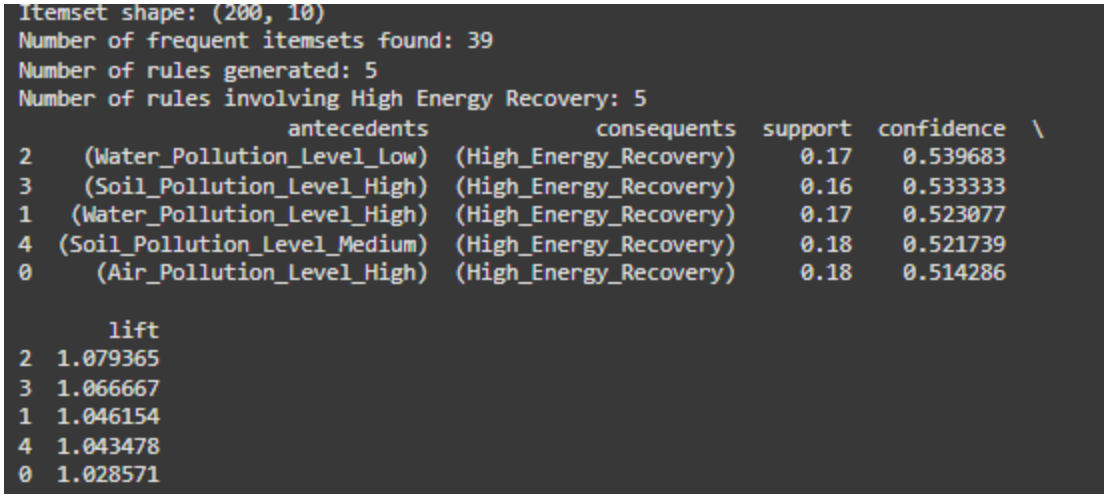
- Integrate external data sources like weather patterns or industrial activity for enriched context.

5.3 Visualization & Monitoring

- Automate the rule mining pipeline to update regularly with new pollution data.

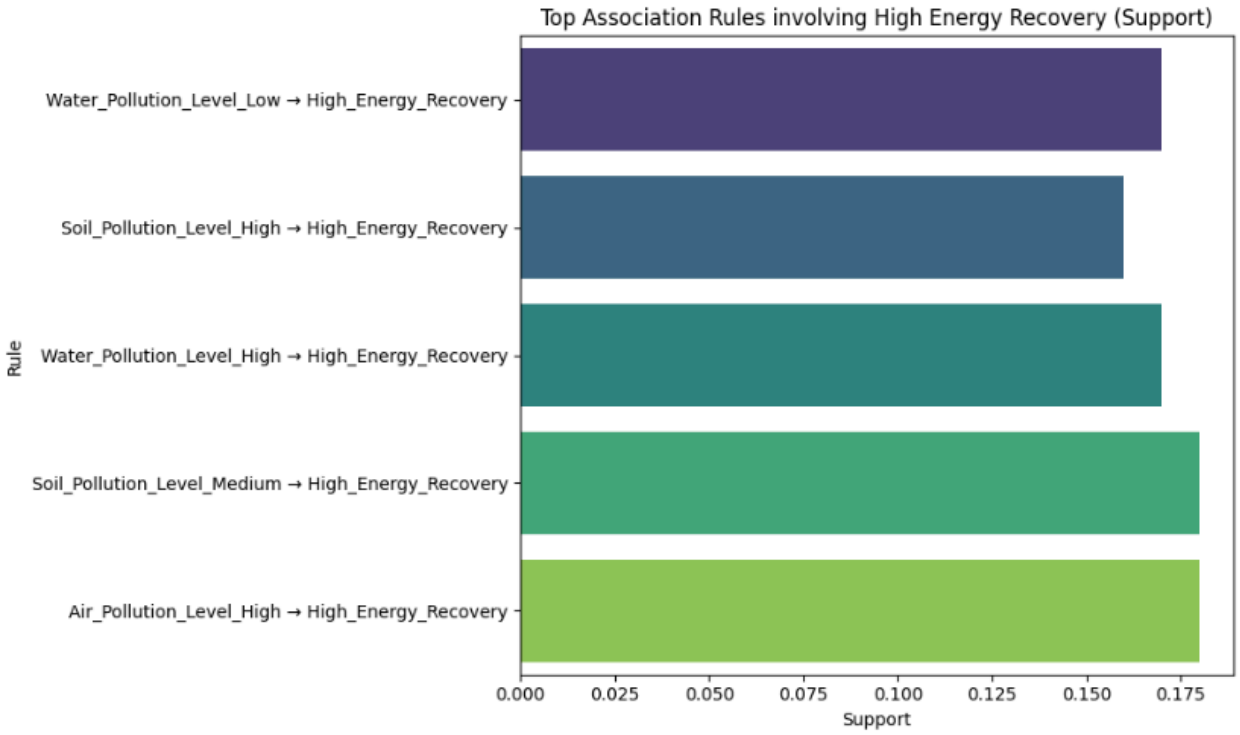
This project showcased the power of combining association rule mining in environmental data science. Apriori identified actionable pollution patterns related to energy recovery.

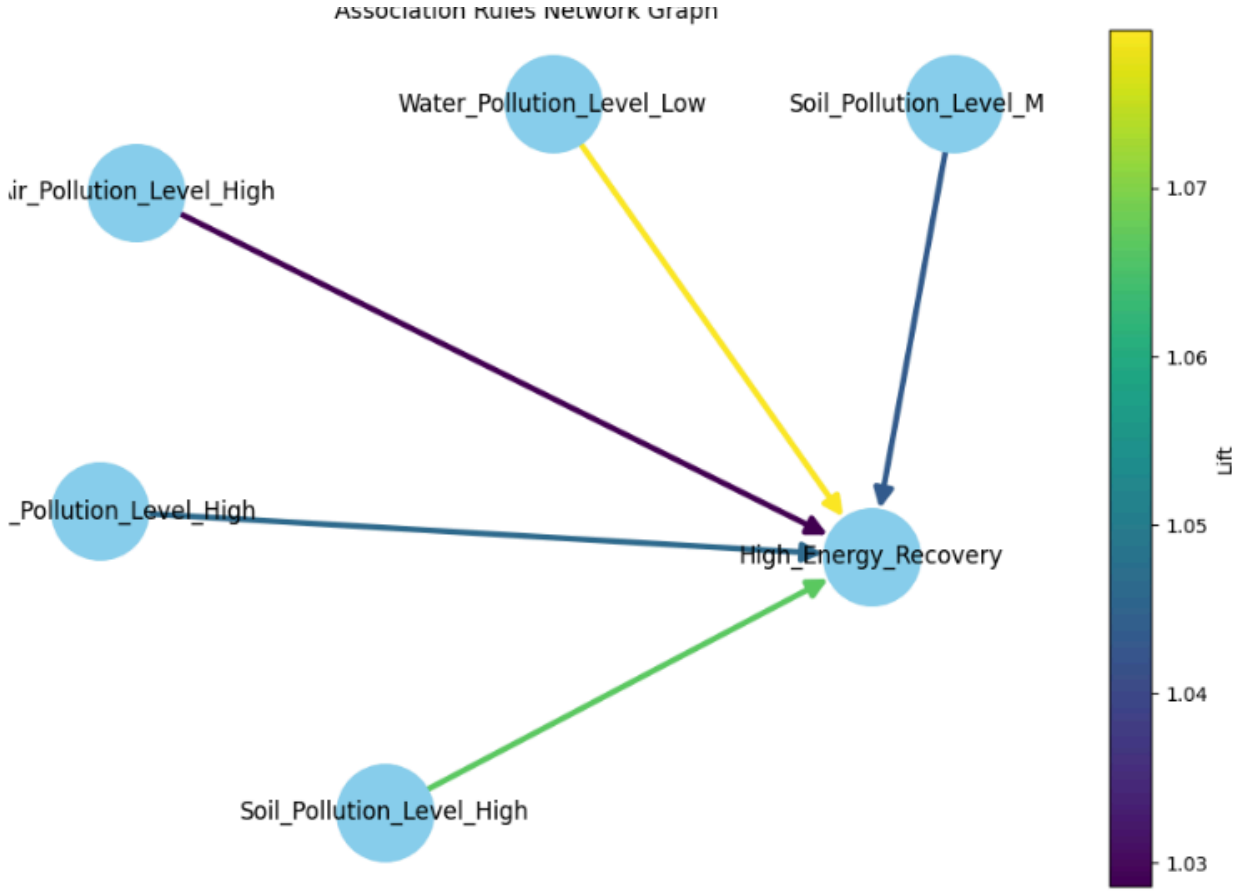
Data Visualizations



```
Itemset shape: (200, 10)
Number of frequent itemsets found: 39
Number of rules generated: 5
Number of rules involving High Energy Recovery: 5
                   antecedents            consequents  support  confidence  \
2    (Water_Pollution_Level_Low)  (High_Energy_Recovery)     0.17    0.539683
3    (Soil_Pollution_Level_High)  (High_Energy_Recovery)     0.16    0.533333
1   (Water_Pollution_Level_High)  (High_Energy_Recovery)     0.17    0.523077
4  (Soil_Pollution_Level_Medium)  (High_Energy_Recovery)     0.18    0.521739
0     (Air_Pollution_Level_High)  (High_Energy_Recovery)     0.18    0.514286

       lift
2  1.079365
3  1.066667
1  1.046154
4  1.043478
0  1.028571
```


Top Association Rules involving High Energy Recovery (Support)
Water_Pollution_Level_Low → High_Energy_Recovery
Soil_Pollution_Level_High → High_Energy_Recovery
Water_Pollution_Level_High → High_Energy_Recovery
Soil_Pollution_Level_Medium → High_Energy_Recovery
Air_Pollution_Level_High → High_Energy_Recovery
Rule
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Support


Association Rules Network Graph
Water_Pollution_Level_Low
Soil_Pollution_Level_M
ir_Pollution_Level_High
_Pollution_Level_High
High_Energy_Recovery
Soil_Pollution_Level_High
1.07
1.06
1.05
1.04
1.03
Lift