Lab 6 Assignment: Association Rule Mining (Apriori Algorithm)

Objectives

1. To apply **data preprocessing techniques** specific to transaction data.
2. To implement the **Apriori Algorithm** to find frequent itemsets.
3. To generate and interpret **Association Rules** using Support, Confidence, and Lift metrics.

Tools and Dataset

- **Tool:** Python with the **`mlxtend`** library (for Apriori) and **`pandas`**.
- **Dataset:** **`online_retail.csv`**
  - **Note:** This dataset requires grouping by **`InvoiceNo`** to treat each invoice as a single transaction before one-hot encoding.

Key Concepts (Review)

- **Support:** The relative frequency of an itemset. (How common is this purchase?)
- **Confidence:** The probability that the consequent is bought, given the antecedent was bought. (How reliable is the rule ?)
- **Lift:** Measures the strength of the association. indicates a strong, positive correlation.

Lab Tasks (Coding Implementation)

Perform the following steps using the **`online_retail.csv`** dataset.

1. **Data Preprocessing and Loading:**
   - Load the **`online_retail.csv`** file.
   - Handle any missing values (NaN) if necessary.
   - **Crucially:** Group the data by **`InvoiceNo`** to convert the DataFrame rows (individual items) into the required **list-of-lists (transaction)** format.
   - Use the **`TransactionEncoder`** from `mlxtend` to one-hot encode the transaction data into a Pandas DataFrame.
2. **Frequent Itemsets Discovery:**
   - Use the **`apriori`** function to find all frequent itemsets. Use a minimum support threshold of **0.03** (3%). *Note: Students may need to adjust this value if their machine struggles with memory or if they get too few/too many results.*

3. **Association Rule Generation:**
    - Use the `association_rules` function to generate all possible rules from the frequent itemsets.
    - Set the evaluation **metric** to `"lift"` with a minimum threshold of **1**.
4. **Rule Inspection:**
    - Sort the generated rules by `lift` (descending) and then by `confidence` (descending).
    - Inspect and print the **top 5 rules**.

Exercises (Analysis and Interpretation)

Use the resulting DataFrame of association rules to answer the following questions:

1. **Highest Confidence Rule:**
    - Find the single rule with the **highest confidence** value.
    - Explain what this high confidence value specifically means for the customer's buying behavior.
2. **High Lift Analysis:**
    - Find any rule where the **Lift is greater than 3**.
    - Explain in simple terms what a high lift value (like) signifies about the relationship between the items, in contrast to a rule with .
3. **Targeted Consequent (Goal-Oriented Rule):**
    - Filter the rules to only show those where the consequent is **"POSTAGE"** (which represents the shipping fee and is common).
    - Sort these specific rules by **Lift** (descending) and state which are the **top 3 Antecedents** that are most strongly associated with a transaction requiring postage.