**TASK2**

**data preparation steps and the transformation for association rule mining based on the provided dataset and code:**

**1. Data Familiarization and Quality Check:**

- The dataset contains 51290 entries and 25 columns.

- Data types include int64, float64, and object (string).

- There are some missing values in the 'Postal.Code' column (9994 non-null out of 51290), which may need handling.

**2. Data Cleaning:**

- Missing values: Since the 'Postal.Code' column has a significant number of missing values, you've chosen to drop rows with missing values using `df.dropna(inplace=True)`.

- Duplicates: You've also removed duplicate rows using `df.drop\_duplicates(inplace=True)`.

**3. Data Transformation for Association Rule Mining:**

- The dataset has been transformed into a transactional format suitable for association rule mining. Each row represents a unique 'Order.ID' with a list of corresponding 'Product.Name' items in the 'Products' column.

- The transformed data is saved to a new CSV file named 'transactions.csv' for association rule mining.

**Package for Association Rule Mining:**

mining algorithms in Python

Overall, the data preparation involves cleaning the dataset by handling missing values and duplicates, and then transforming it into a suitable format for association rule mining, ensuring that each transaction is represented properly for mining frequent itemsets and generating association rules.

**Task1(b)**

1. **Support Level**:

- Support measures the frequency of itemset occurrence in the dataset.

- A lower support threshold may yield more itemsets but may include less meaningful or rare associations.

- A higher support threshold may result in fewer but more significant itemsets.

For this analysis, a **minimum support of 0.01 (1%)** is chosen. This level strikes a balance between capturing frequent itemsets without overwhelming the analysis with too many rules.

2. **Confidence Level**:

- Confidence measures the reliability or strength of the association rule.

- A lower confidence threshold may result in more rules but could include weaker associations.

- A higher confidence threshold ensures stronger and more reliable rules but may filter out potentially interesting but less strong associations.

For this analysis, a **minimum confidence of 0.3 (30%)** is selected. This level ensures that the discovered rules are reasonably strong and reliable without being too restrictive.

Now, let's analyze one interesting association rule from the output provided:

- **Rule**: Accessories → Phones

- **Antecedent** (If): Accessories

- **Consequent** (Then): Phones

- **Support**: 8.8953% (0.088953)

- Confidence: 100.00% (1.0)

- Lift: 5.410937

- **Kulczynski** Coefficient: 54.45% (0.544477)

- Imbalance Ratio: 51.87% (0.518679)

**Contingency** **Table**:

|  |  |  |
| --- | --- | --- |
|  | Accessories (Antecedent) | Not Accessories (Antecedent) |
| Phones (Consequent) | 8.8953% (0.088953) | 91.1047% (0.911047) |
| Not Phones (Consequent) | 0% (0.0) | 91.1047% (0.911047) |

**Analysis**:

1. **Support**: The rule occurs in nearly 8.8953% of transactions, indicating a moderate occurrence frequency.

2. **Confidence**: If Accessories are in a transaction, it always includes Phones, showing a strong association.

3. **Lift**: Accessories increase the likelihood of also including Phones by a factor of 5.410937.

4. **Kulczynski Coefficient**: Indicates a moderate balance between support and confidence, supporting the rule's reliability.

5. **Imbalance Ratio**: The rule's occurrence is moderately imbalanced, suggesting a notable disparity between Accessories and Phones.

**Justification**:

- **Interesting Rule**: This rule is interesting because it shows a strong and reliable association between Accessories and Phones. Such insights can be valuable for targeted marketing or product placement strategies.

- **Support and Confidence Levels**: The chosen support and confidence levels balance capturing meaningful associations while ensuring rule reliability and relevance to the business context.

Task1 (c )

1. Multilevel **Frequent** **Itemsets**:

- The frequent itemsets include combinations of Category and Sub-Category, indicating their co-occurrence patterns in the dataset.

- For example, `(**Category\_3**, **Sub.Category\_3**)` has a support of 8.8953%, suggesting a notable occurrence of products in this category and sub-category combination.

**2. Multilevel Association Rules:**

- The discovered rule `(Category\_1) ⇒ (Sub.Category\_3)` has a confidence of 48.13% and a lift of 3.51.

- This rule indicates that if a product is in Category\_1, there's a 48.13% chance it belongs to Sub.Category\_3 as well, which is 3.51 times more likely than random chance.

- Similarly, the reverse rule `(Sub.Category\_3) ⇒ (Category\_1)` has a higher confidence of 64.81%, indicating a stronger association from sub-category to category.

**Justification for Interesting Rules:**

- **Redundancy Check**: The rules are not redundant because they capture distinct associations between categories and sub-categories, providing insights into how products are grouped and sold together.

- **Interesting Insights**: These rules are interesting as they reveal patterns of product categorization and sub-categorization, which can be valuable for inventory management, cross-selling strategies, and understanding customer preferences.