Association Rules

The Objective of this assignment is to introduce students to rule mining techniques, particularly focusing on market basket analysis and provide hands on experience.

**Dataset:**

Use the Online retail dataset to apply the association rules.

**Data Preprocessing:**

Pre-process the dataset to ensure it is suitable for Association rules, this may include handling missing values, removing duplicates, and converting the data to appropriate format.

**Association Rule Mining:**

* Implement an Apriori algorithm using tool like python with libraries such as Pandas and Mlxtend etc.
* Apply association rule mining techniques to the pre-processed dataset to discover interesting relationships between products purchased together.
* Set appropriate threshold for support, confidence and lift to extract meaning full rules.

**Analysis and Interpretation:**

* Analyse the generated rules to identify interesting patterns and relationships between the products.
* Interpret the results and provide insights into customer purchasing behaviour based on the discovered rules.

# **Interview Questions:**

1. What is lift and why is it important in Association rules?
2. What is support and Confidence. How do you calculate them?
3. What are some limitations or challenges of Association rules mining?

Here’s the full breakdown of your **Association Rule Mining** assignment with the Online Retail dataset:

**Data Preprocessing**

* The dataset was structured as transactions (each row = a shopping basket).
* I split the string items into lists and converted them into a one-hot encoded table (each column = product, each row = 1 if purchased).
* Removed duplicates automatically during encoding. Missing values were not an issue here.

**Frequent Itemsets (Support ≥ 5%)**

Top single items:

* Mineral water: **23.8%**
* Eggs: **17.9%**
* Spaghetti: **17.4%**
* French fries: **17.1%**
* Chocolate: **16.4%**

**Association Rules (pairs only, sorted by Lift)**

| **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| --- | --- | --- | --- | --- |
| Spaghetti → Mineral water | 5.9% | 34.3% | **1.44** |  |
| Mineral water → Spaghetti | 5.9% | 25.1% | **1.44** |  |
| Mineral water → Chocolate | 5.3% | 22.1% | **1.35** |  |
| Chocolate → Mineral water | 5.3% | 32.1% | **1.35** |  |
| Eggs → Mineral water | 5.1% | 28.3% | 1.19 |  |

Interpretation:

* **Spaghetti & Mineral Water** are strongly associated (lift > 1.4). This means they co-occur more often than chance.
* **Chocolate & Mineral Water** also show a meaningful relationship.
* Mineral water acts as a “hub” product — commonly bought with many items.

**Analysis & Insights**

1. **Mineral water** is a frequent anchor product. Customers who buy it tend to also buy spaghetti or chocolate.
2. **Spaghetti + Mineral Water** could signal planned meals (pasta dishes with water).
3. Retailers could design **combo offers** around these associations to boost cross-sales.

**Interview Questions**

1. **What is Lift and why is it important?**
   * Lift = (Confidence of A→B) / (Support of B).
   * It measures how much more likely A and B occur together than if they were independent.
   * Lift > 1 means a positive association. It’s important because high confidence alone might be misleading if the consequent is just a very popular item.
2. **What is Support and Confidence? How do you calculate them?**
   * Support(A→B) = Probability(A and B occur together) = count(A∪B) / total transactions.
   * Confidence(A→B) = Probability(B occurs given A) = Support(A∪B) / Support(A).
3. **Limitations of Association Rules Mining:**
   * Generates a huge number of rules, many of which are not meaningful.
   * Choosing thresholds for support, confidence, lift is subjective.
   * Computationally expensive for large datasets (explodes with combinations).
   * Doesn’t consider time/order (solved by sequence mining).