# Assignment 3

▷▷ PROBLEM STATEMENT: Apply a-priori algorithm to find frequently occurring items from given data and generate strong association rules using support and confidence thresholds.

▷▷ OBJECTIVE: Model associations between products by determining sets of items frequently purchased together, and building association rules to derive recommendations

▷▷ OUTCOMES: Demonstrated market basket analysis using a-priori algorithm to find frequently occurring items from given data and generate strong association rules using support and confidence thresholds.

▷▷ HARDWARE & SOFTWARE REQUIREMENTS: Any CPU (i3 processor or higher), 4GB RAM or higher, 1 GB HDD or 128GBSSD minimum; 64 bit UNIX/LINUX OS, Python3, jupyter, apyori

▷▷ THEORY:

Association Rule Mining finds interesting associations and relationships between large sets of data items. This rules show~~ing~~s how frequently an itemset occurs in a transaction. It is defined as an implication expressio of the form X→Y, where X, Y are any 2 itemsets.

Market Basket Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy to together frequently, so that items can be strategicaly placed next to each other to boost sales.

Some definitions:

↳ Support count: frequency of itemset occurence

↳ Frequent itemset: itemset whose support >= minsup threshold.

The Association Rule Evaluation Metrics are:

1) support (S): The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.

2) confidence (c): It is the ratio of the number of transactions that includes all items in an itemset {A} ~~to~~ ~~the~~ as well as the transactions that include all items in {A} to the number of transactions That includes all items in {A}.

3) Lift (l): The lift of the rule X → Y is the confidence of the rule divided by the expected confidence, assuming ~~the~~ expected ~~confidence~~ ~~such~~ that itemsets {X}, {Y} are independent of each other. The expected confidence is the confidence divided by the frequency of {Y}. ~~to~~ Itemsets occur together: i) as expected if lift = 1,

ii) more than expected if lift >1, and iii) less than expected if lift <1.

APRIORI ALGORITHM :

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases, first proposed by Agrawal and Srikant in 1994.

It is designed to operate on databases containing transactions, with each transaction being a set of items (itemsets).

Given a threshold C, the algorithm identifies the item sets which are subsets of at least C transactions in the database.

It uses a 'bottom up' approach, where frequent subsets are extended one at a time (called candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no other successful extensions are found.

A-priori uses breadth first search and a hash tree to count candidate item sets efficiently. It generates k item sets of length k-1; then it prunes the candidates which have infrequent sub pattern. After that, it scans the transaction database to determine frequent item sets among the candidates.

## ▷▷ CONCLUSION:

Thus frequently occurring items from given Market Basket dataset and strong association rules using support and confidence thresholds were found using a-priori algorithm.