**Business Intelligence Lab**

**Experiment 09**

**Aim:**

To implement Association mining using RapidMiner and Python

**Theory:**

**Association Mining:**

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of the dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of machine learning, and it is employed in Market Basket analysis, Web usage mining, continuous production, etc. Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

Market Based 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 together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

**TID Items**

1 Bread, Milk

2 Bread, Diaper, Beer, Eggs

3 Milk, Diaper, Beer, Coke

4 Bread, Milk, Diaper, Beer

5 Bread, Milk, Diaper, Coke

Before we start defining the rule, let us first see the basic definitions.

**Support Count** – Frequency of occurrence of a itemset.

Here Support count({Milk, Bread, Diaper})=2

**Frequent Itemset** – An itemset whose support is greater than or equal to minsup threshold.

**Association Rule** – An implication expression of the form X -> Y, where X and Y are any 2 itemsets.

Example: {Milk, Diaper}->{Beer}

**Rule Evaluation Metrics –**

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.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

Support = support count(X+Y) / total –

It is interpreted as a fraction of transactions that contain both X and Y.

Confidence(c) –

It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.

Conf(X=>Y) = Supp(XUY) / Supp(X) –

It measures how often each item in Y appears in transactions that contain items in X also.

Lift(l) –

The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.

Lift(X=>Y) = Conf(X=>Y) \div Supp(Y) –

Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.

**Types of Association Rule Learning**

Association rule learning can be divided into three algorithms:

1. Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

2. Eclat Algorithm

The Eclat algorithm stands for Equivalence Class Transformation. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster than the Apriori Algorithm.

3. F-P Growth Algorithm

The F-P growth algorithm stands for Frequent Pattern, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

**Applications of Association Rule Learning**

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

1. Market Basket Analysis: It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
2. Medical Diagnosis: With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
3. Protein Sequence: The association rules help in determining the synthesis of artificial Proteins.
4. It is also used for Catalog Design and Loss-leader Analysis and many more other applications.

**Implementing Association Mining using Python Libraries**