IntelliPaat

# Data Science

## Association Rule Mining

## &

## Recommendation Engine


Copyright IntelliPaat. All rights reserved

# Agenda

01 Association Rule Mining

02 Apriori Algorithm

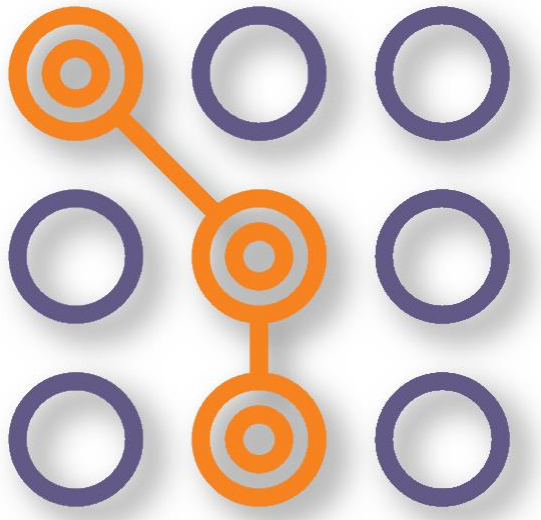
03 Apriori Algorithm in R

Copyright IntelliPaat. All rights reserved

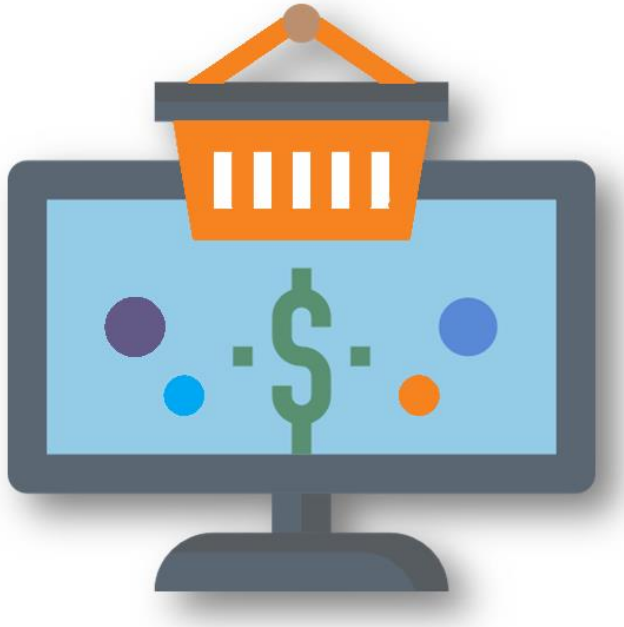
# Association Rule Mining

Copyright IntelliPaat. All rights reserved

# Association Rule Mining



Association rules analysis is a technique to uncover how items are associated to each other

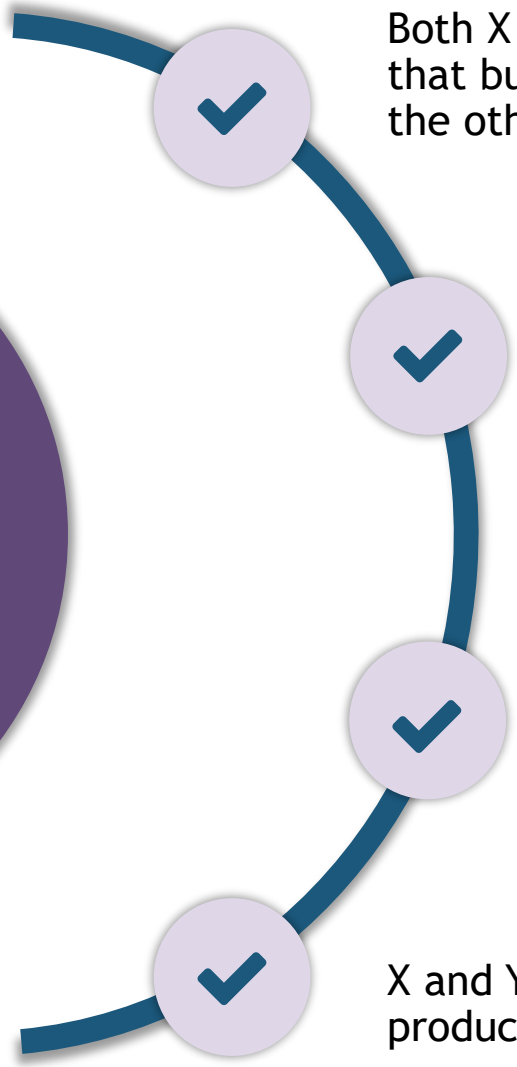


Association mining is usually done on transactions data from a retail market or from an online ecommerce store

Copyright IntelliPaat. All rights reserved

# Association Rule Mining

Understanding customer's **buying patterns** can help to increase sales in several ways. If there is a pair of items, X and Y, that are frequently bought together:



- Both X and Y can be placed on the same shelf, so that buyers of one item would be prompted to buy the other
- Promotional discounts could be applied to just one out of the two items
- Advertisements on X could be targeted at buyers who purchase Y
- X and Y could be combined into a new product, such as having Y in flavors of X

Copyright IntelliPaat. All rights reserved

# Association Rule Mining

1. Since most transactions data is large,
the Apriori Algorithm makes it easier to find these patterns or *rules* quickly

2. So, what is a *rule*?
A rule is a notation that represents which item/s is frequently bought with what item/s

3. It has an LHS and an RHS part and represented as follows:

   itemset A => itemset B

4. This means, the item/s on the right were frequently purchased along with items on the left

Copyright IntelliPaat. All rights reserved

# Measures of Association Rule Mining

Copyright IntelliPaat. All rights reserved

# Measures

There are three common ways to measure association

**Support**

**Confidence**

**Lift**

Copyright IntelliPaat. All rights reserved

# Support

This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears

$$Support = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} = P(A \cap B)$$

Copyright IntelliPaat. All rights reserved

# Support

It is an important measure because a rule that has very low support may occur simply by chance

A low support rule is also likely to be uninteresting from a business perspective because it may not be profitable to promote items that customers rarely buy together

For these reasons, Support is often used to eliminate uninteresting rules

Copyright IntelliPaat. All rights reserved

# Confidence

This says how likely item B is purchased when item A is purchased, expressed as {A -> B}

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P(A \cap B)}{P(A)}$$

Copyright IntelliPaat. All rights reserved

# Confidence

Confidence measures the reliability of the inference made by a rule

This is measured by the proportion of transactions with item A, in which item B also appears

For a given rule A → B, the higher the confidence, the more likely it is for B to be present in transactions that contain A. Confidence also provides an estimate of the conditional probability of B given A

Copyright IntelliPaat. All rights reserved

# Lift

Lift says how likely item B is purchased when item A is purchased, while controlling for how popular item B is

$$ExpectedConfidence = \frac{Number\ of\ transactions\ with\ B}{Total\ number\ of\ transactions} = P(B)$$

$$Lift = \frac{Confidence}{Expected\ Confidence} = \frac{P(A \cap B)}{P(A) \,.\, P(B)}$$

Copyright IntelliPaat. All rights reserved

# Lift

***Lift*** is the factor by which the co-occurrence of A and B exceeds the expected probability of A and B co-occurring, had they been independent

So, higher the lift, higher the chance of A and B occurring together

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm

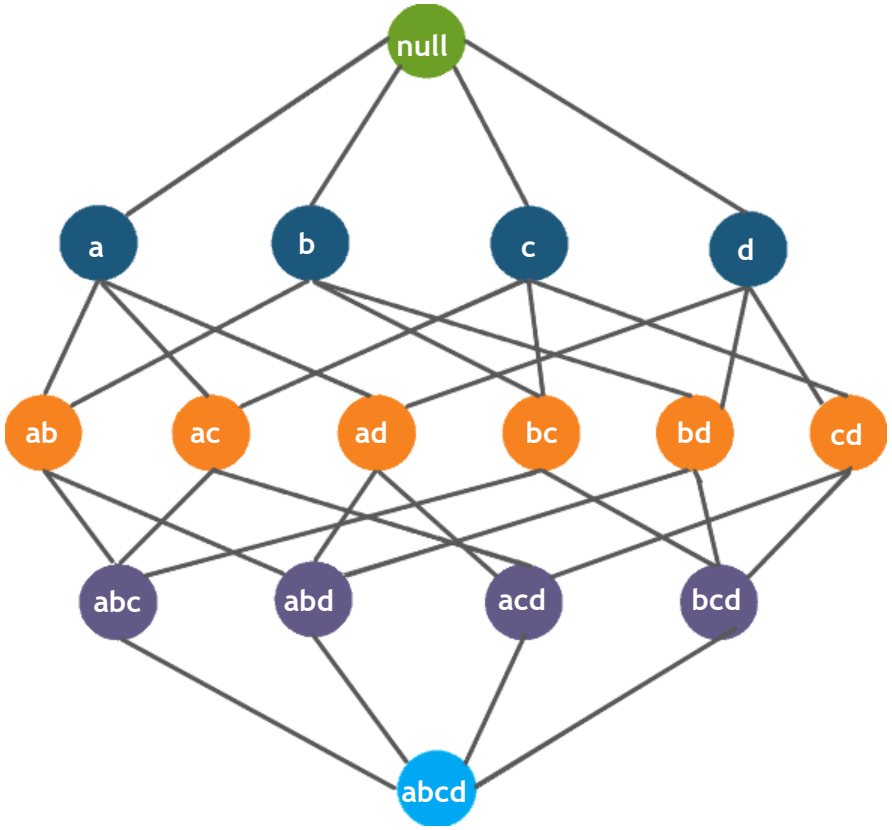
Copyright IntelliPaat. All rights reserved

# Apriori Algorithm

If an itemset is frequent, then all of its subsets must also be frequent. A key concept in Apriori algorithm is the anti-monotonicity of the support measure

It assumes that:

- All subsets of a frequent itemset must be frequent
- Similarly, for any infrequent itemset, all its supersets must be infrequent too




Copyright IntelliPaat. All rights reserved

# Working of Apriori Algorithm

The entire algorithm can be divided into two steps:

**Step 1**

***Frequent itemset generation***

- Find all itemsets for which the support is greater than the threshold support following the process we have already seen earlier in this article

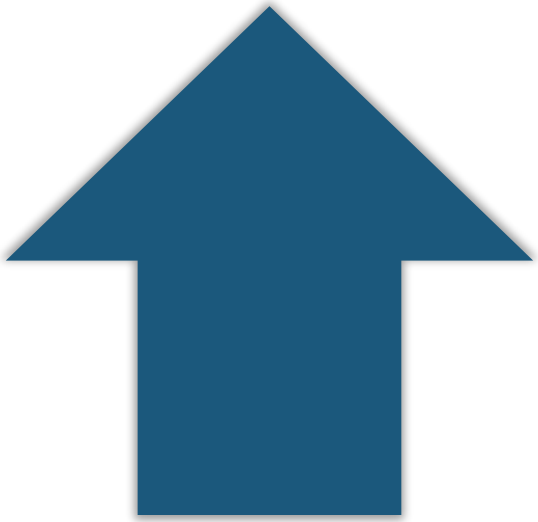
**Step 2**

**Rule generation**

- Create rules for each frequent itemset using the binary partition of frequent item-sets and look for the ones with high confidence. These rules are called candidate rules
- For finding association rules, we need to find all rules having support greater than the threshold support and confidence greater than the threshold confidence

Copyright IntelliPaat. All rights reserved

# Pros & Cons of Apriori Algorithm



**Pros**

- It is an easy-to-implement and easy-to-understand algorithm
- It can be used on large itemsets

**Cons**

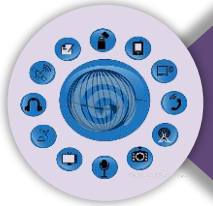
- Sometimes, we may need to find a large number of candidate rules which can be computationally expensive
- Calculating support is also expensive because it has to go through the entire database

Copyright IntelliPaat. All rights reserved

# Association Rule Mining Applications

Copyright IntelliPaat. All rights reserved

# Association Rule Mining Applications



Identifying value-added services on mobile subscription and recommending what products can be offered to a customer (parallelism)

Understanding which symptoms tend to co-morbid to help improve patient care and medicine prescription

Understanding customer transition so that they can tie up with other vendors for additional discounts

Planning efficient public services and helping public businesses for setting up new factories, shopping malls and so on, as well as marketing particular products

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R

Building the Apriori algorithm on top of the “Groceries” dataset

```
> summary(Groceries)
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda           yogurt          (Other)
            2513             1903             1809             1715             1372            34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29   14   14    9
  21   22   23   24   26   27   28   29   32
  11    4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
       labels  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

Copyright IntelliPaat. All rights reserved

# Tasks to be performed

1. Build an Apriori rule on the “Groceries” dataset where the support value is “0.002” & the confidence value is “0.5”

2. Build an apriori rule on the “Groceries” dataset where the support value is “0.007” & the confidence value is “0.6”

Copyright IntelliPaat. All rights reserved

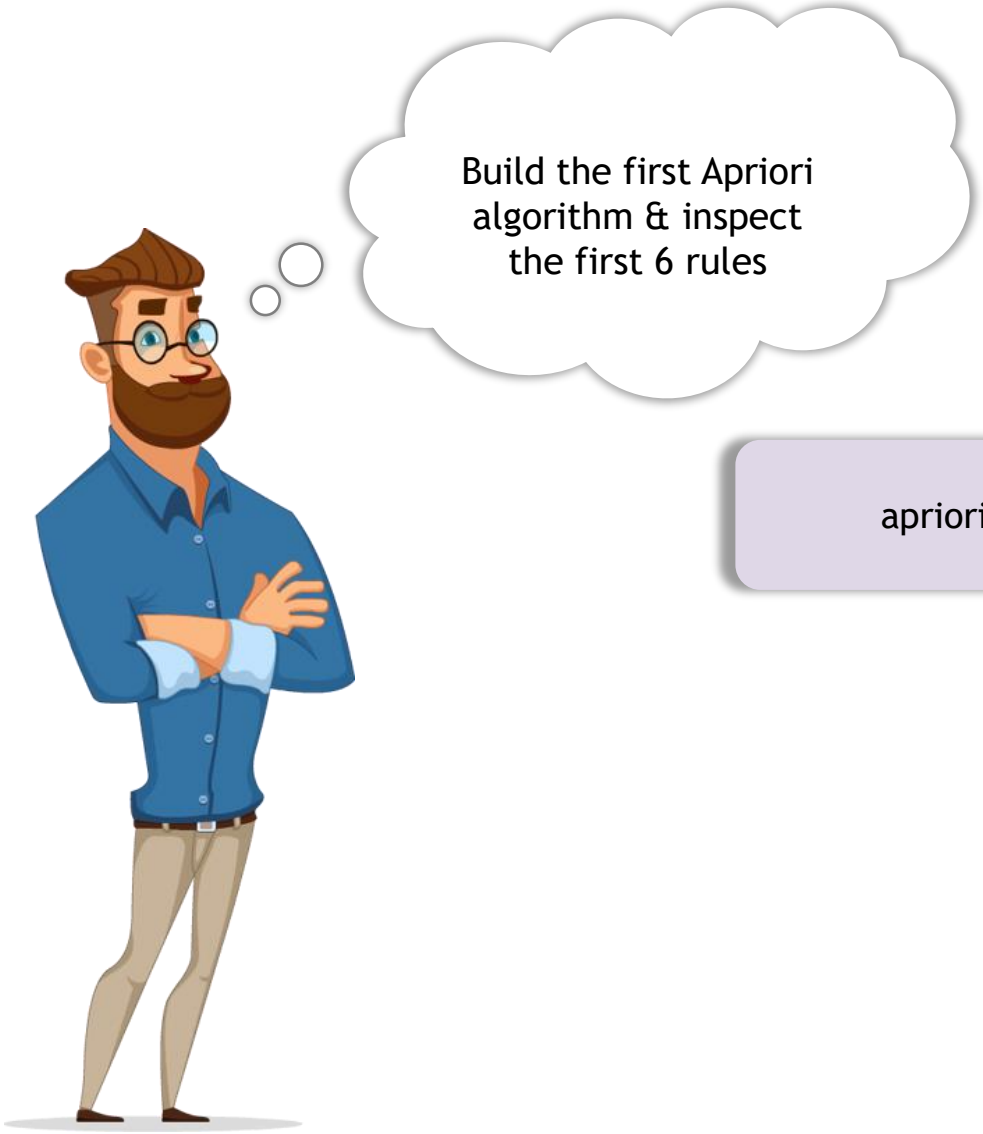
# Apriori Algorithm in R


```
library(arules)

library(arulesviz)

data(“Groceries”)
```

```
summary(Groceries)
```

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R




```
apriori(Groceries,parameter = list(support=0.002, confidence=0.5)) -> rule1
```

```
inspect(head(rule1))
```

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R


```
inspect(head(sort(rule1,by="lift"),5))
```

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R


```
plot(rule1)

plot(rule1,method="grouped")
```

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R


```
apriori(Groceries,parameter = list(support=0.007, confidence=0.6)) -> rule2
```

```
inspect(head(rule2),4)
```

Copyright IntelliPaat. All rights reserved

# Apriori Algorithm in R


```
plot(rule2)

plot(rule2,method="grouped")
```

Copyright IntelliPaat. All rights reserved

# Quiz

Copyright IntelliPaat, All rights reserved

# Quiz

Which package would you have to load to work with the 'apriori()' function?

1. Association
2. Apriori
3. Arules
4. Support

Copyright IntelliPaat. All rights reserved


# Thank You

Copyright IntelliPaat. All rights reserved

India : +91-7847955955

US : 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com

24X7 Chat with our Course Advisor

Copyright IntelliPaat. All rights reserved