A

Project Report

on

MARKET BASKET ANALYSIS

Submitted for the partial fulfilment

of B.Tech. Degree

in

COMPUTER SCIENCE & ENGINEERING

by

Shubham Gupta (1805210052)
Shivam Sachan (1805210048)
Ossama Bin Hassan (1805210032)

Under the supervision of

Prof. D.S. Yadav

Ms. Deepa Verma

Department of Computer Science and Engineering

Institute of Engineering and Technology

Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh.

Contents

Declaration

We accept that this submission is our own work and, within our knowledge and beliefs, for materials previously published or created by others, or for the awarding of a degree. We hereby declare that we do not include any material or material errors or diplomas from the university or other universities. However, if the recognition is not stated in the text. I didn't submit my project to another institution to request another degree.

Submitted by: - Date: 25/05/2022

(1) Name: Shubham Gupta

Roll No.: 1805210052

Branch: CSE

Signature:

(2) Name: Shivam Sachan

Roll No.: 1805210048

Branch: CSE

Signature:

(3) Name: Ossama Bin Hassan

Roll No.: 1805210032

Branch: CSE

Signature:

Certificate

This is a project report titled "**Market Basket Analysis**" submitted by Shubham Gupta, Shivam Sachan, and Ossama Bin Hassan for the awarding of a Bachelor of Engineering degree in Computer Science and Engineering, is a record of work they did under my supervision and guidance at the Computer Science and Engineering Department at Institute of Engineering and Technology, Lucknow.

It has also been proven that, as far as we know, this project has not yet been submitted to other institutions to confer other degrees.

Prof. D.S Yadav

Department of Computer Science and Engineering

Institute of Engineering and Technology, Lucknow

Ms. Deepa Verma

Department of Computer Science and Engineering

Institute of Engineering and Technology, Lucknow

Acknowledgement

We are deeply grateful to our Head of Department Prof. D.S. Yadav and our supervisors, Prof. D.S. Yadav, Ms. Deepa Verma and Dr Alok Misra of the Institute of Engineering and Technology, Lucknow, India for their guidance, patience and support. We consider ourself very fortunate for being able to work with a very considerate and encouraging mentors like them. Without their offering to accomplish this project, we would not be able to finish our project successfully.

We would also like to thank our fellow students for their insightful comments and constructive suggestions for improving the quality of this project work.

Name: Shubham Gupta

Roll No.: 1805210052

Signature:

Name: Shivam Sachan

Roll No.: 1805210048

Signature:

Name: Ossama Bin Hassan

Roll No.: 1805210032

Signature:

Abstract

Market basket analysis is one of the main methods retailers use to increase sales by better understanding their customers' buying behavior. A large amount of data is analyzed. B. Identify purchase history, product groups, and products that can be purchased together. It works by looking at the combination of things that happen together on a regular basis on sale. Organizational rules are often used in the analysis of shopping carts and transaction data and aim to identify strict rules derived from transaction data using interest rates based on the concept of strict rules. So in this project we have taken a data set of about 5 Lakh items from different clients that are processed and sorted to include the database into a usable format written in another csv file such as MBA.csv. After receiving the data in the required format the frequency section is checked to obtain the most purchased items. Now the apriori algorithm is used to find organizational rules, general rules can be issued and special laws can also be issued such as milk and bread bought together with great confidence. In the last step we remove all the obsolete rules that are followed by visualizing the rules that have been developed into different elements. Finally important rules and decisions can be drawn from graphs and tables.

It turns out that the high affinity is not surprising, as the taste of one is bought by another from the same family. As mentioned earlier, one of the most common uses of mining organization rules is in the area of program schedules. Once a pair of items is determined to be relevant, you can make recommendations to your customers to increase sales. And in the process, we hope to be able to introduce our customers to things they haven't tried or thought they existed.

List of Figures

List of Symbols & Acronyms

MBA : Market Basket Analysis

SKU : Stock Keeping Unit

FP : Frequent Pattern

OCVR : Overall Variability of association rules

LHS: Left hand side

RHS: Right hand side

Chapter-1

INTRODUCTION

Market basket analysis is a way to delve into the data that merchants use to understand customers behavior better so they can enhance their sales. Market basket analysis is a well-known and widely used method used by large retailers to emphasize relationships between products such as bread and butter. It works by looking for product combinations that sometimes appear in the transaction. To see it from a different perspective, ask the trader to look at the relationships between individual purchases. With the continued growth of information technology, companies are collecting and storing large amounts of data. This includes analysis of large datasets such as B. Purchase history, product group identification, and products that can be bought together.

The spread of shopping cart analysis is being driven by the advent of electronic point-of-sale (POS) systems. Compared to many records held by store owners, the digital records generated by the PS system make it easier for businesses to process and analyze large amounts of purchase data.

Implementing market basket analysis requires a background in mathematics and data science in addition to computer programming skills with specific algorithms. For those who do not have the necessary technical skills, there are off-the-shelf trading tools available.

Companies need to turn this data into useful information and information in order to make decisions in changing markets. Knowing that a customer who buys one product is likely to buy another, the seller sells those products together or makes the consumer more optimistic about the second product. can. If a customer who buys a diaper might buy beer, the beer may be on display right next to the diaper corridor. For young fathers, even if they can get what they and their children need before the weekend begins, it's an immediately satisfying problem. The strength of such a relationship is not important and can be used for sale or sale.

The dataset is obtainded by the open source Instacart. The purchasing process is as follows: At first, the user orders groceries from the app. The local consumer will receive the notification of the order, go to the nearest shop, buy items and deliver them to the customer. Data set data for over 3 million orders from over 200,000 users. Records are flexible depending on the order and order time. Therefore, time and order related characteristics were created to predict whether the product will be redesigned.

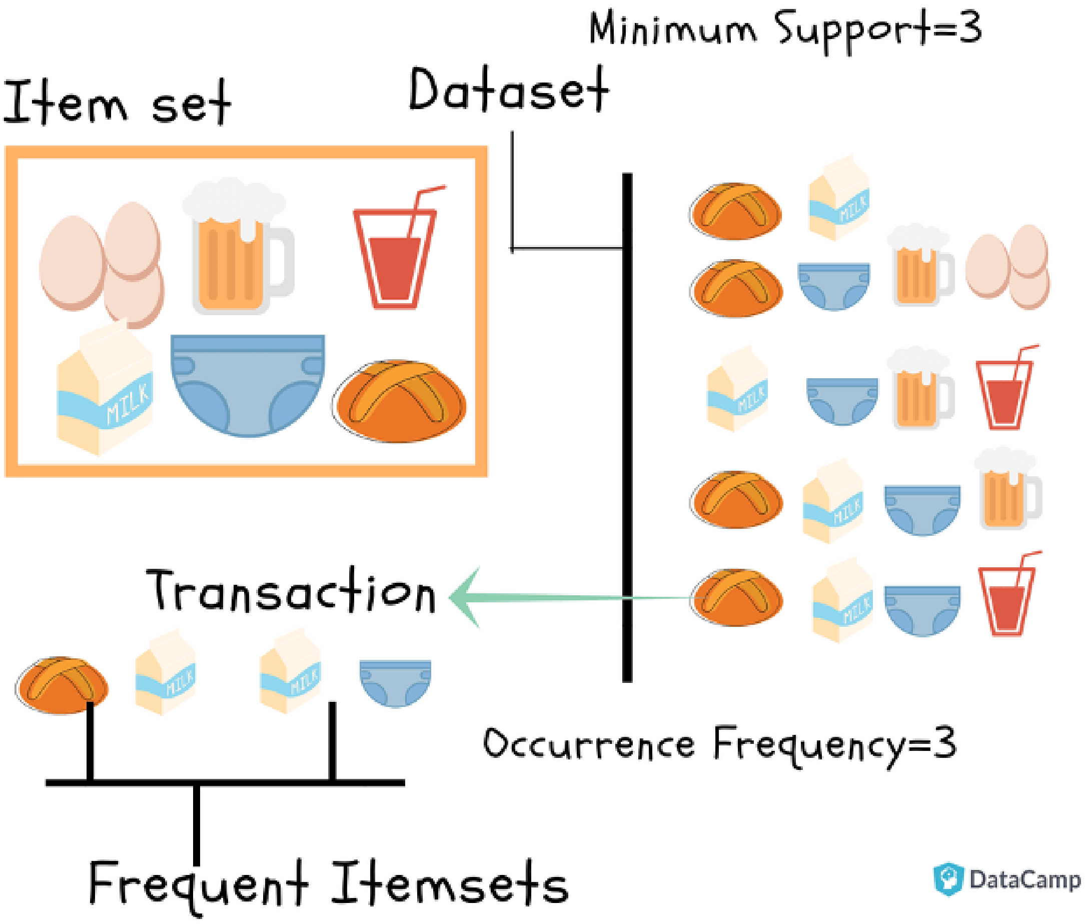

Minimum Support=3
Item set
Dataset
MILK
Transaction
Occurrence Frequency=3
Frequent Itemsets
DataCamp

Figure 1

Chapter-2

LITERATURE REVIEW

Agrawal et.al., 2013 Market basket analysis is a practical subject rather than an academic subject, so most of the research on this subject is done in real retail stores. MBA is an old field of data mining and one of the best examples of mining association rules.

Roshan Gangurde , 2017 He also concluded that major retailers can use MBA to attract more customers, increase the value of their shopping carts, and carry out more profitable advertising and promotions. This study also proposed designing and developing intelligent predictive models to generate mapping rules that can be applied to recommender systems to make features more operational. Then, in late 2017, we designed a method optimized for MBA with the goal of predicting and analyzing consumer buying behavior. In this study, we introduced a new algorithm based on data cleaning, which is one of the most important challenges in all areas of data analysis. To solve this challenge, they combined two data mining algorithms: a priori algorithms and neural networks. They also emphasized that one of the biggest challenges is that customer demands are constantly changing in terms of seasons and time. Also, MBA issuance varies depending on the season and season, so you will have to redo it many times.

Solnet, et.al.,2016 They explored the possibility of market basket analysis to increase hotel revenue in 2016. Researchers have researched and derived the most attractive services and products that can attract and satisfy hotel guests and encourage them to make repeated purchases. This approach allows you to increase sales without increasing the number of customers.

Ngai, et.al,2009 They announced that data mining in customer relationship management is a new trend that helps identify, attract, retain, and develop customers. Customer retention and development is important for maintaining long-term customer relationships. Data mining promises many opportunities in the market sector. Method classification, grouping, and regression are most commonly used to identify customers. Custom development typically uses methods such as classification, regression, association detection, pattern recognition, and optimization.

Kamley et.al., 2014 They developed an association rule mining model in 2014 to find interesting patterns in stock market datasets. This model helps predict stock prices and helps stock brokers and investors to understand market conditions and invest in the right direction. June 2015, S.S. Nandgaonkar mining association rules for forecasting stock markets from Umbarkar and S.S. financial news. The valuation depends on the technical trading indicators and the closing price of the stock.

Kapadia, In 2015, He conducted a survey to analyze consumer behavior of lifestyle store products. It gives valuable insights into the formation of baskets. This survey helped with product assortment, inventory management of potentially sold items, inducing price increases for potentially sold items, offering discounts to loyal customers, and cross-selling. rice field. The limitation of this study was that its scope was limited to one store in a particular area.

Mohammed Al-Maolegi, 2014 From the International Journal on Natural Language Computing from February 2014, Mohammed Al Maolegi concludes that the improved Apriori is improved by reducing the time it takes to scan the submissions of the candidate item set by reducing the number of transmissions to scan. Whenever the kofkitemset increases, the gap between our improved Apriori and original Apriori increases from the view of time. Will be consumed. It takes less time to generate a candidate surplus with the improved weaving than with the original weaving. The improved Ariori reduces the time required by 67.38%.

Mohini H. Chandwani, 2018 Market basket analysis is perfect for making business decisions. B. What is for sale, how it is put on the shelves to maximize profits, etc. This analysis performs an analysis of the spread transfer data. However, advances in barcode technology have made it easy to save data and collect large amounts of data. These records are typically stored in tertiary storage due to limited database functionality.

Gupta et.al, 2014 In today's business, most companies have branches in different areas. To maintain the sales economy of these stores. The chain is getting bigger and bigger. Wal-Mart, for example, is the largest supermarket chain in the world. The discovery of purchase patterns at these multiple stores changes both in time and place. Basic association rules are not valid in this multi-store chain. 13. Author: Hussein and Hussein, Year: 2012 Hussein and Hussein (2012) We used a data mining approach that included market basket analysis with data on student attendance. This analysis helps identify groups of students with nearly identical absenteeism. This type of similarity may emphasize that students are absent from class due to peer pressure and is not an acceptable reason. This technique was tested by analyzing attendance data from more than 2,000 students who attended a public higher education institution for the first semester. The results obtained helped me find students who were absent from class just because their friends were absent from class.

After reading research papers mentioned above, we got a brief knowledge of how we can perform market basket analysis and we found two algorithms, namely Apriori and FP-Growth algorithms. We will be using Apriori algorithm as this algorithm uses candidate generation where frequent item sets are extended one item at a time.

Market Basket Analysis

Market basket analysis is a method of identifying the strength of the association between a pair of products purchased together and a pattern of opportunities. One cause is when two or more things happen together.

Market basket analysis creates If-then scenario rules. Eg, if item A was purchased, item B could be purchased. The rules are stochastic. In other words, it is derived from the frequency of observations. Frequency is the percentage of baskets that contain items of interest. This rule can be used in pricing strategies, product placements, and different types of cross-selling strategies.

Working of market basket Analysis

For the sake of clarity, consider a market basket analysis related to shopping in a supermarket. Shopping cart analysis gets transaction-level data that lists all the items that a customer has purchased in a single purchase. Technology determines the relationship between purchased products and other products. Then use these relationships to create a profile that contains If Then rules for purchased items.
The rule can be written as follows:
If {A} Then {B} The “IF” part of the

rule is known as the antecedent, and the “THEN” part of the rule is the consequent. The prefix is a condition and the result is the result. Association rules have three majors that represent the confidence of the rule: Surrort, Confidence, and Lift. For example, you are in a supermarket to buy milk. According to the analysis, are you more likely to buy the same shape of apples or cheese than those who didn't buy the milk?

Another example of Association Rules

Lets suppose, there are 100 customers

10 of them purchased bread,

8 purchased jam and

6 purchased both.

purchased bread => purchased jam

Support = P(Bread & Jam) = 6/100 = 0.06

Confidence = support/P(Jam) = 0.06/0.08 = 0.75

Lift = confidence/P(Bread) = 0.75/0.10 = 7.5

Some Important Terms:

● **Support** is the relative frequency indicated by the rule. In many cases, it's a good idea to look for a high level of support to make sure it's a useful relationship. However, a little help can be helpful when trying to find a "hidden" relationship.

This is the default frequency for items. Mathematically, the surrogate for item A is just the ratio of transactions containing A to the total number of transactions.

Surrort (Grares) = (Transactions including Greres) / (Total transaction) Surrrort (Grares) = 0.666

● **Confidence** is a measure of rule reliability. If the confidence in the above example was half of the cases where bread and jam were bought, the purchase also included butter and milk. 50% Confidence can be claimed as a strong recommendation, but in moderate situations this level may not be high enough. It is the likelihood that customers who bought both A and B. It divides the number of transactions involving both A and B by the number of transactions involving B.

Confidence (A => B)

= (Transactions involving both A and B)/(Transactions involving only A) Confidence({Grapes, Apple}=> {Mango})

= Support(Grapes, Apple, Mango)/Support(Grapes, Apple)

= 2/6 / 3/6

= 0.667

● **Lift** is the ratio of the observed support to that expected if the two rules were independent.

As a basic rule of thumb, a lift value close to 1 means that the rules are completely independent. In general, a lift value> 1 is more "interesting" and shows some useful rule patterns.

This period denotes that there is an enhancement in the sale of X when you sell Y.

Lift(X => Y) = Confidence(X, Y) / Support(Y).

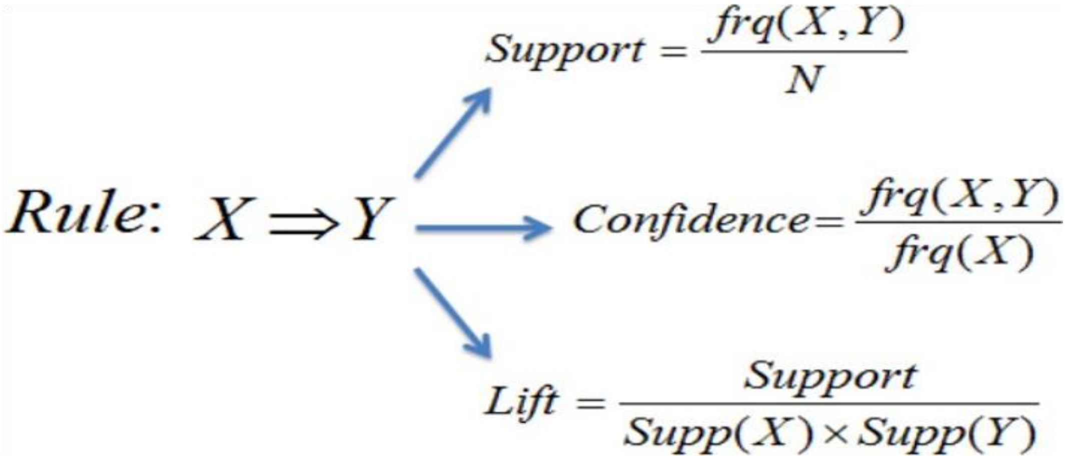
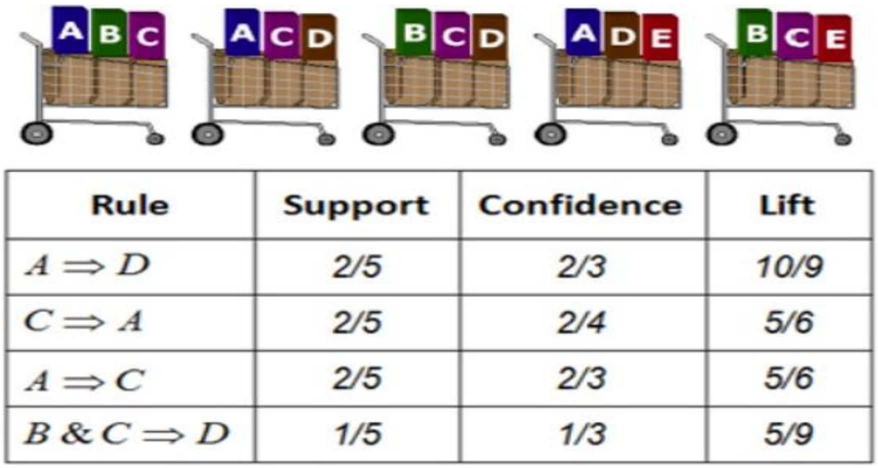

$Rule: X \Rightarrow Y$
$Support = \frac{frq(X,Y)}{N}$
$Confidence = \frac{frq(X,Y)}{frq(X)}$
$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

Figure 2


A B C
A C D
B C D
A D E
B C E

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Figure 3

Lift ({Grapes, Mango} => {Apple}) = 1

So, the probability that a customer will buy both X and Y together is many times higher than simply buying them.

1. If Lift = 1 then there is no relationship among the itemset.
2. If Lift > 1 then products in the itemset X, and Y, are more likely to be purchased together.
3. If Lift < 1 then products in itemset X, and Y, are unlikely to be purchased together.

Practical Applications of Market Basket Analysis:

1. **Retails**: In retail, market basket analysis helps determine which items are bought together, in sequence, and seasonally.

2. **Banks**: In finance, you can use shopping cart analysis to analyze a customer's credit card purchases for fraud detection and cross-selling opportunities.

3. **Medical**: In the healthcare or medical field, market basket analysis can be used for pathological and symptom analysis. It can be used to better identify the disease profile. Some Data requirements for this analysis are Basket(Column identifying an individual basket) and Products(Items included in each basket).

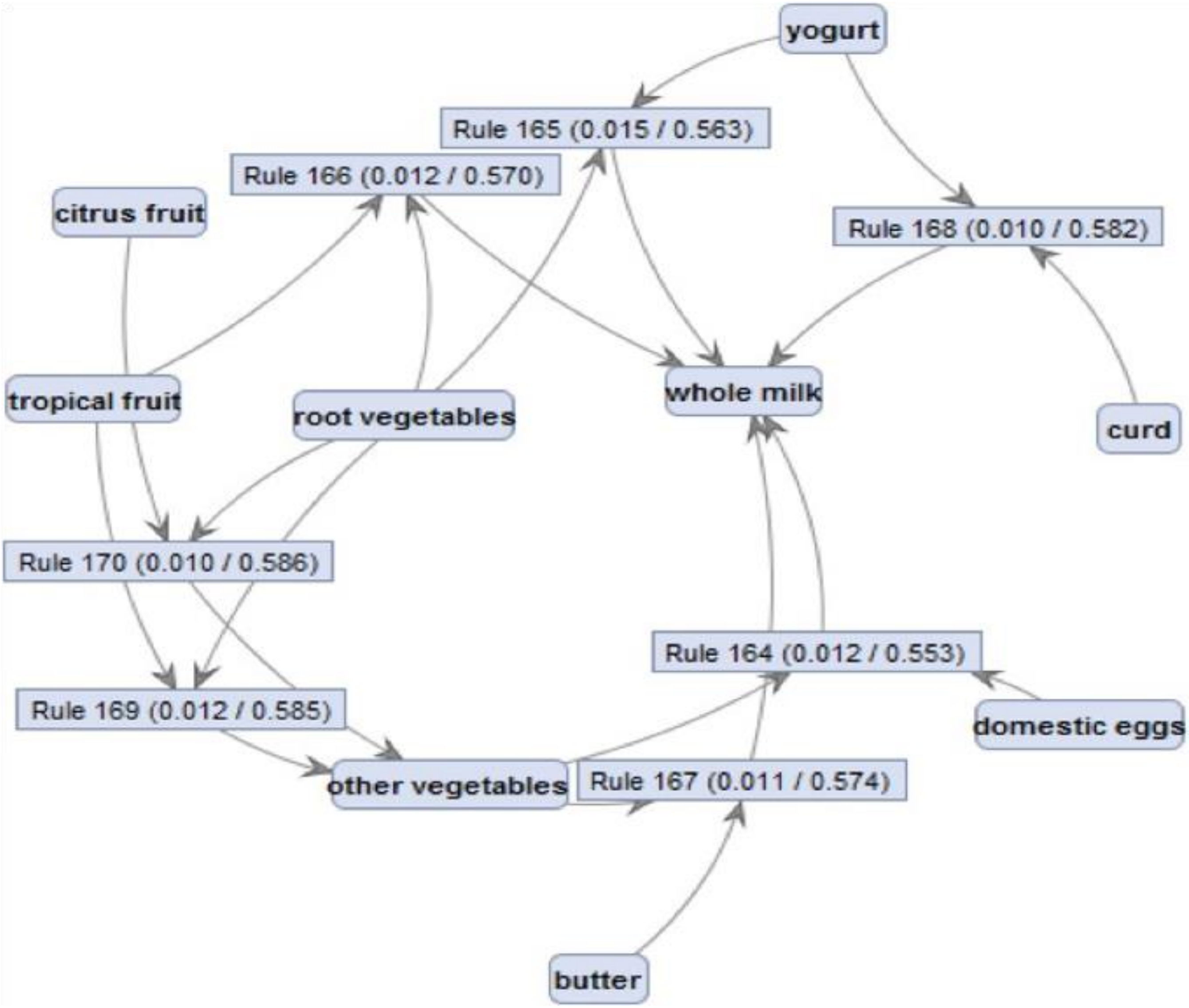


Figure 4

Types of Market Basket Analysis (MBA):

1. **Descriptive MBA**: This type derives insights only from residual data and is the most commonly used approach. The analysis here is not a judgment, it just uses statistical techniques to evaluate the agreement between the products. For those familiar with the basics of data analysis, this type of modeling is known as unsupervised learning.

2. **Predictive MBA**: This type considers items purchased in sequence to determine to cross-sell. his type uses supervised learning models such as classification and regression. It is essentially intended to mimic the market to analyze what can happen. For example, purchasing an extended warranty is more likely after you purchase your iPhone. Basically, consider the items purchased in order to determine cross-selling. It's not as widely used as the descriptive MBA, but it's still a very valuable tool for marketers.

3. **Differential MBA**: This type takes into account data from various shops and buys made by distinct customer groups at separate times. If the rule applies to one dimension and not the other, analysts can determine the factors that are causing the exception. This type of analysis is a useful tool for analyzing competitors. Compare buy records between stores, seasons, two periods, different days of the week, and more to find relationship patterns of customer behavior.

So in our project, we aim to use predictive market basket analysis. Major algorithms already working in this domain are based on the association are AIS, SETM, Apriori.

About the Dataset:

We have used the dataset from UCI Machine Learning Repository which has details of about 5 Lakh orders. A good Dataset is an asset for a nice working project which requires preprocessing which is explained below.

Dataset Pre-processing:

Similar to some other algorithms for general item set mining. B. Ariori, Eclat, Fgrowth checks the transition database as follows: The first scan determines the frequency of individual element sets (for individual elements) of the element. Rare items cannot be part of a common set of items and are therefore removed from the transaction. In addition, the items in each transaction are sorted in descending order in relation to a particular recommender system and frequency in the database. The algorithm does not depend on this particular order, but experiments have shown that it runs much faster than a random order. In my experiments, ascending order leads to particularly slow operations, which is even worse than random ordering. (In this respect, FPgrowth works the exact opposite of Arriori, which tends to run fastest when items are sorted in ascending order, but is the same as Esclat. This is because the item is hidden from the customer. You will also benefit from falling into the pattern that was done.

The OCVR (Overall variability of association rules) enables further development and understanding of association rules. The rules for getting high scores in the OCVR index may be of special interest. These rules appear to be systematically and significantly changing. That is, consumer buying behavior is less stable and non predefined in this particular basket.

You need to monitor and control the rules with the highest level of trust first. By adopting the right marketing strategy, you are more likely to raise the level of trust in these rules to a higher level. This seems to be achieved randomly and sometimes occasionally throughout the year.

At the initial step we have changed the date format to suitable format and invoice number to numeric format, followed by checking for all the empty spaces. Later when we don’t want certain columns we have set them as NULL as we don’t want dates in our algorithm for mining.

Chapter-3

METHODOLOGY

The goal of recommender systems is to create meaningful recommendations for articles and products that are of interest to your collection. Basic algorithms such as B. Arriori and Fr Growth gather knowledge about people's tastes and recognize that people who buy spaghetti and wine are generally more interested in sauces. Association rules are an important part of the development and recommendation engine. Association rules override multiple rules when a date set containing information from a spread basket expires. Each rule contains a collection of product names as predecessors, the resulting product names, and several class measures:

REQUIREMENT ANALYSIS

1) Software Requirements:
   - a. RStudio
   - b. Web Browser
   - c. R
   - d. Excel
2) Hardware Requirements:
   - a. i5 processor/ Ryzen 3 processor above
   - b. 2GB GPU Card
   - c. 4GB RAM or Above
3) Libraries
   - a. Arules
   - b. ArulesViz
   - c. Plyr
   - d. Dplyr
   - e. Tidyverse
   - f. Readxl
   - g. Ggplot2
   - h. Lubridate

3.1 COLLECTION OF DATA

The dataset used for this project is a specific fileset that shows the customer's orders over time. This is anonymous user data and contains a selection from many users of the UCI Machine Learning Archive and a food record of 5 lakh. For eachl customer, UCI provide between the range of 1 and 100 of their orders, placement of products purchased together for each order, the order was placed in, and the sizeable time between orders to specify.

3.1.1 Dataset Understanding

Aisles

The dataset contains 8 columns such as Country, Date of purchase, Invoice Number etc.

Departments

This dataset has the following departments. All department names are listed in alphabetically ordered below. ‘Frozen’, ‘other’, ‘pantry’, ‘meat seafood’, ‘babies’, ‘canned goods’, ’ ‘bakery’, ‘produce’, ‘household’, ‘breakfast’, ‘alcohol’.

Products

Within 8 columns and many departments, there are about 10,000 items in the file.

Orders

Let's examine the user configuration. For example, if a user had previously placed 100 orders, the last order was a toy train. Note that order number 1 is meaningless for the day as it is the first order with no previous record.

3.2 PREPROCESSING DATASET

3.2.1 Working with empty values

A technique called assignment was used to replace zeros with non-zero values. Let's calculate the total number of zeros in each column of the dataset. The first step is to use .isnull () to see which cells in the dataframe are null, then use an aggregate function to count the number of nulls in each column, we use .isnull.sum ().

3.2.2 Replace the number in the day field with a text value

Since VeryFirstOrder was left blank in the original record, replace the days_since_prior_order field with the null value FirstOrder. The day of the week I purchased was listed in the number 18 on the original record, so I've replaced the number with a text value for clarity.

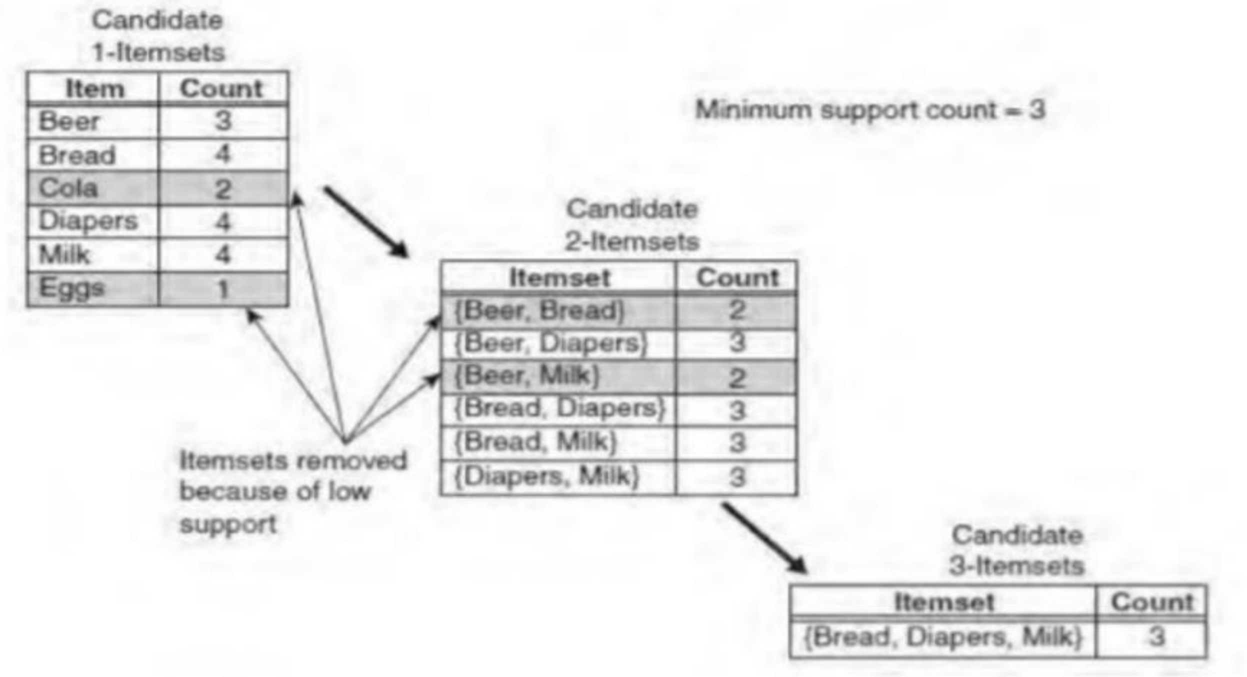
3.3 APRIORI ALGORITHM

The Apriori algorithm is a fundamental algorithm developed by Agrawal & Srikant in 1994 to find the relationship between the frequent item sets of Boolean assignment rules. Apriori's principle states that "if the item set is common, then all of its subset gadgets may be common." If the item set support is greater than the super level, the item set is "common". This name comes from the term "previous". The Ariori algorithm contains association rules of type in data mining. The thumb rule that shows the relationships between many attributes is often called affinity analysis or MBA.

For example, to understand Apriori's principles, consider whether it is an item set {b, d, e} from a dataset that is a common item set. 25) Then all subsets such as b, d, e, {b, d}, {b, e}, {d, e} are also a common set of items. As a result, if {b, d, e} are common, all subtypes b, d, e must be regular. In contrast, if elemental units like {a, b} are uncommon, then all supersets need to be even rarer. You can completely trim the entire segment, including the {a, b} superset. The method of cutting the direction of the road according to the height of the roadbed is called overlapping cut. The nature of runaway is prevented by an important goal of the support scale. This Sharasteristis, also known as anti-monotonicity, indicates a flaw in support measures.

Apriori algorithm working:

1. Create all common itemsets – A more occurring itemset is a set of objects that has transaction support greater than minimal assist.
2. Create all assured association policies from more occurring itemsets – A assured association rule is a rule with confidence greater than minimum confidence. To use this algorithm on the dataset, the Apriori class is taken out and this is inherited from the Apyori library.



Minimum support count = 3
Itemsets removed because of low support

Candidate 1-Itemsets

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

Candidate 2-Itemsets

Itemset	Count
{Beer, Bread}	2
{Beer, Diapers}	3
{Beer, Milk}	2
{Bread, Diapers}	3
{Bread, Milk}	3
{Diapers, Milk}	3

Candidate 3-Itemsets

Itemset	Count
{Bread, Diapers, Milk}	3

Figure 11: Illustration On Apriori Algorithm

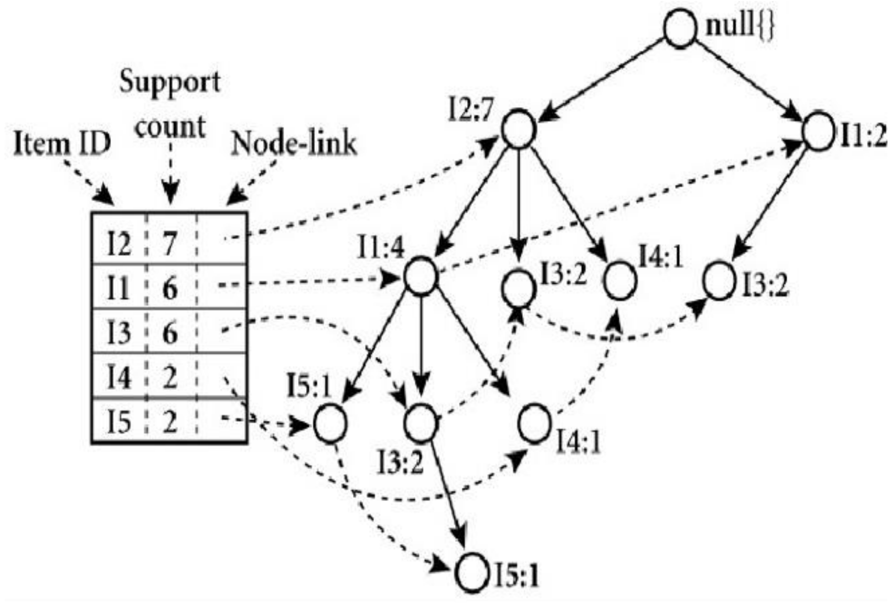
Figure 5

In the above case, a different article is previously accepted as a 1-point candidate. Items set below the minimum support were discarded after we relied on their support. As a result, {Eggs} & {Cola} are excluded. The next iteration uses a common 1itemset to generate candidate 2itemsets. A wide variety of possible 2 item sets generated using this algorithm is 6 (4C2). This is because the 39 has only four common 1 item sets. The brutal strategy of listing all item sets (up to 3 in length) as candidates returns 41 candidates. (6C1) + (6C2) + (6C3) = 6 + 15 + 20 = 41 Considering Ariori's theory, variety is reduced to 13. (6C1) + (4C2) + 1 = 13. This depicts a 68% pruning even in the easy case of the various possible item sets. Pseudocode shown in the algorithm for generating a set of common elements in the Apriori algorithm.

FREQUENT PATTERN GROWTH ALGORITHM

The FP-Growth algorithm provides an alternative way to measure common Item collections by comparing datasets using a tree graph data structure.

Assume FP-Growth Tree as FP and convert the dataset to a graphical representation. Instead of the generation and validation method used in both the algorithm, first creates a structure tree & uses this comparison tree to generate a regular set of items. The efficiency of the FP-Growth algorithm relies on the amount of compression that the can perform while building the tree. FP-Growth has solved the problem of iterating underage searches and then combining suffixes within the detection of various models. Using a slightly retrograde object as a suffix gives the abundant effect. This error significantly reduces search costs.
A typical pattern tree is a tree structure built using an earlier set of items of
data. The main purpose of the FP tree is to mine the most common patterns. Each
node in the FP tree represents an object in this item set. The root node represents the null value and the subordinate nodes represent the item set of data. The mapping
to the subordinate nodes between the item sets of these nodes is maintained while the tree is being built.



Item ID	Support count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	

Figure 6

3.4 ANALYSIS

3.4.1 Analysis of Apriori Algorithm-

Here, at the beginning of the comparison process, we will introduce a pre-algorithm for performance analysis. Apriori is the most effective algorithm for mining item sets. The main goal of Apriori is to create multiple paths through a dataset or database where tests or data are stored. The Apriori algorithm relies on the Apriori property, which states: "A collection of non-empty element sets must be present frequently. Breadth-first search (BFS) is used in the Apriori algorithm. It also uses the downstream blocking property (a better collection of abnormally broken items is abnormal).

Transactional databases are usually placed horizontally. Item set frequency is measured in each transaction.

Limitation of Apriori Algorithm

Often, the algorithm also scans the database, degrading overall performance, so the era of frequent itemsets is the most computationally intensive step. This algorithm presumes that the database is continuous in memory.

In addition, the temporal & spatial complexity of apriori algorithm is very big, O (2 ^ {| D |}), so it will be exponential in nature. Where | D | is the horizontal width (the sum of the object types) that exists in the database.

Optimising Apriori algorithm

Optimize your current a priori algorithm and use the following techniques to save time and memory.

● **Hash primary-based item set count**: k itemset with the corresponding hash bucket issue on the edge is uncommon.

● **Transaction reduction**: Transactions that do not contain the general k itemsets will not be useful in subsequent scans.

● **Partitioning:** A set of objects that are definitely common in a DataBase must be common in one or more than one of the DB's partitions.

● **Sampling:** Extracting on a subset of a particular dataset, the possibility of determining auxiliary threshold lower bound + completeness.

● **Dynamic Itemset Count:** The easiest way to add a new candidate item set while all subsets are estimated to be common.

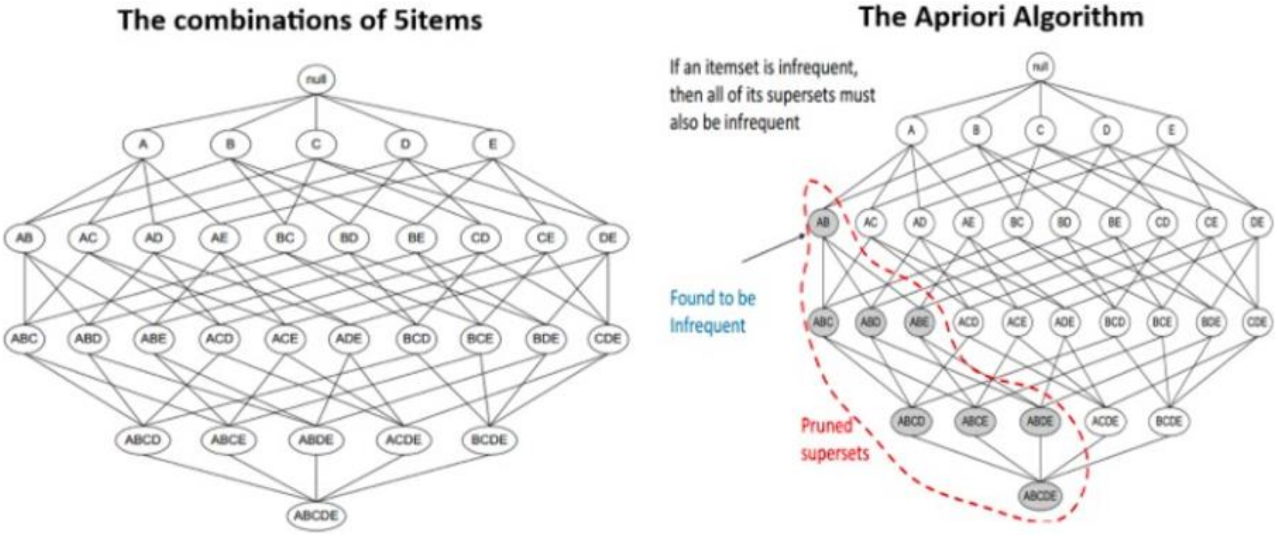
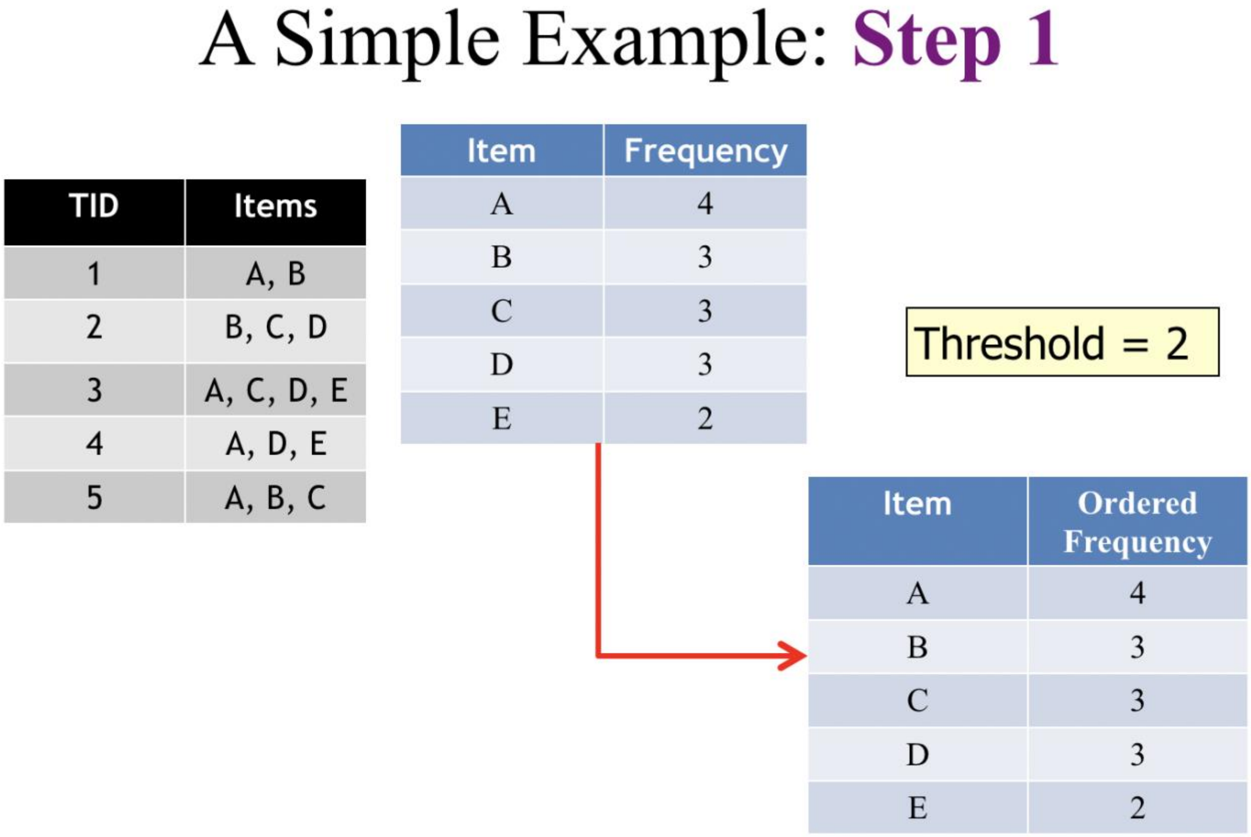


Figure 7

3.4.2 Analysis of FP-Growth algorithm

As far as growth is concerned, this algorithm uses fragmentation and detection techniques and uses the structure of the data to obtain a simplified transmission database. Candidates do not need a regular item set. The FP tree is mined instead of the usual pattern. In the early stages of FP development, a list is created and organized in the order of development of the exchange. A structure called a node. Excluding the root node, different FP nodes contain an element name, a support number, and a pointer connected to a node of a tree with the same element name.



A Simple Example: Step 1

TID	Items
1	A, B
2	B, C, D
3	A, C, D, E
4	A, D, E
5	A, B, C

Item	Frequency
A	4
B	3
C	3
D	3
E	2

Threshold = 2

Item	Ordered Frequency
A	4
B	3
C	3
D	3
E	2

Figure 8

Chapter-4

EXPERIMENTAL RESULTS

Results and Plots-

The frequency plot is shown below which represents the item frequency of different items in absolute manner like lunch bag suki design was brought about 500-1000 times.

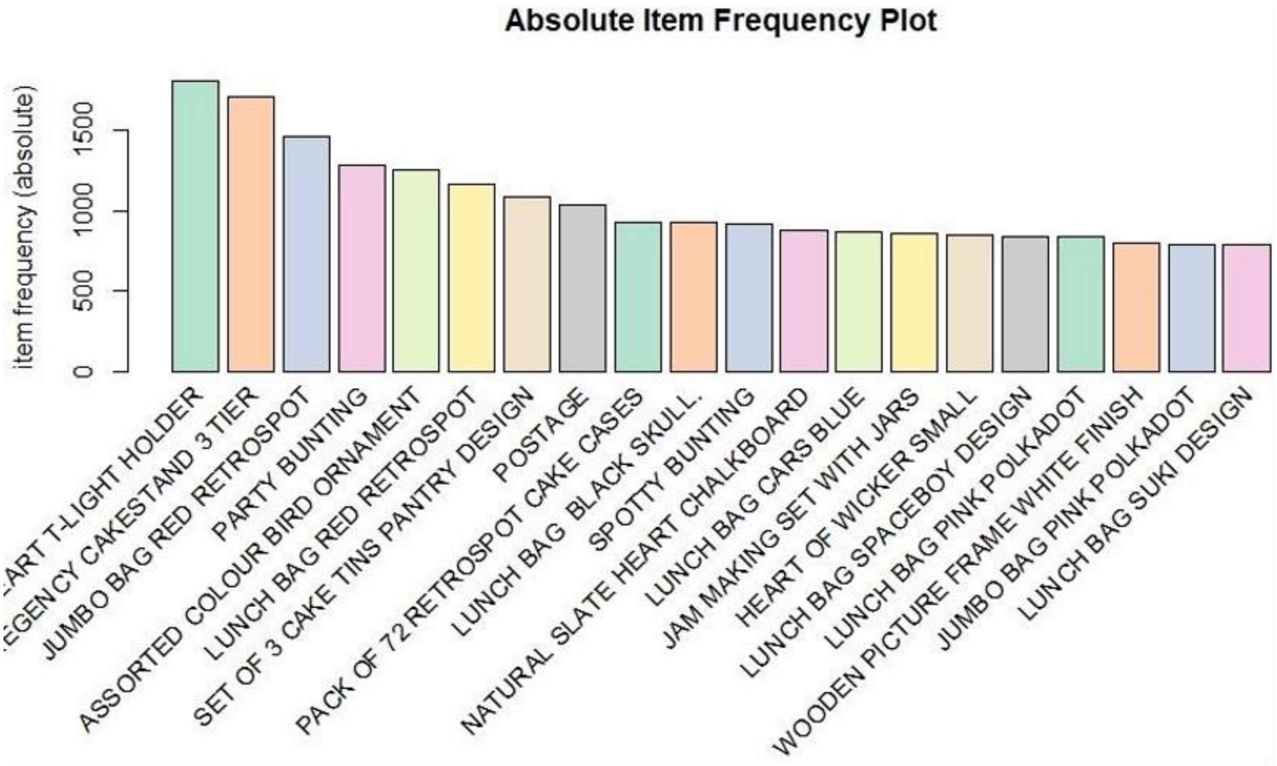
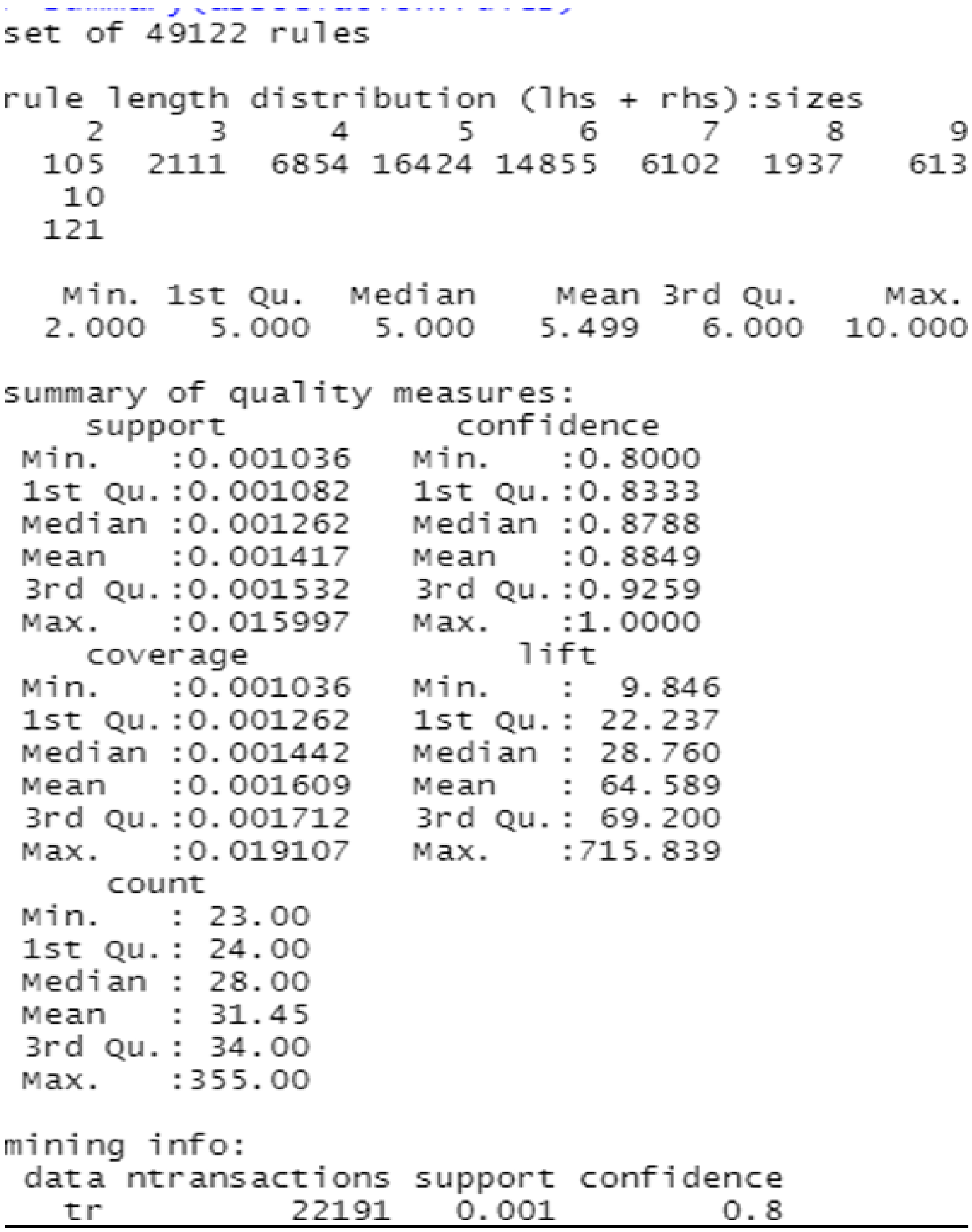


Figure 9

Total rules generated with confidence 0.001 is shown below in figure where around 49000 rules were generated.



```
set of 49122 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9 
  105  2111  6854 16424 14855  6102  1937   613 
   10 
  121 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   5.000   5.000   5.499   6.000  10.000 

summary of quality measures:
    support           confidence    
 Min.   :0.001036   Min.   :0.8000  
 1st Qu.:0.001082   1st Qu.:0.8333  
 Median :0.001262   Median :0.8788  
 Mean   :0.001417   Mean   :0.8849  
 3rd Qu.:0.001532   3rd Qu.:0.9259  
 Max.   :0.015997   Max.   :1.0000  
    coverage             lift        
 Min.   :0.001036   Min.   :  9.846  
 1st Qu.:0.001262   1st Qu.: 22.237  
 Median :0.001442   Median : 28.760  
 Mean   :0.001609   Mean   : 64.589  
 3rd Qu.:0.001712   3rd Qu.: 69.200  
 Max.   :0.019107   Max.   :715.839  
     count       
 Min.   : 23.00  
 1st Qu.: 24.00  
 Median : 28.00  
 Mean   : 31.45  
 3rd Qu.: 34.00  
 Max.   :355.00  

mining info:
 data ntransactions support confidence
   tr         22191   0.001        0.8
```

Figure 10

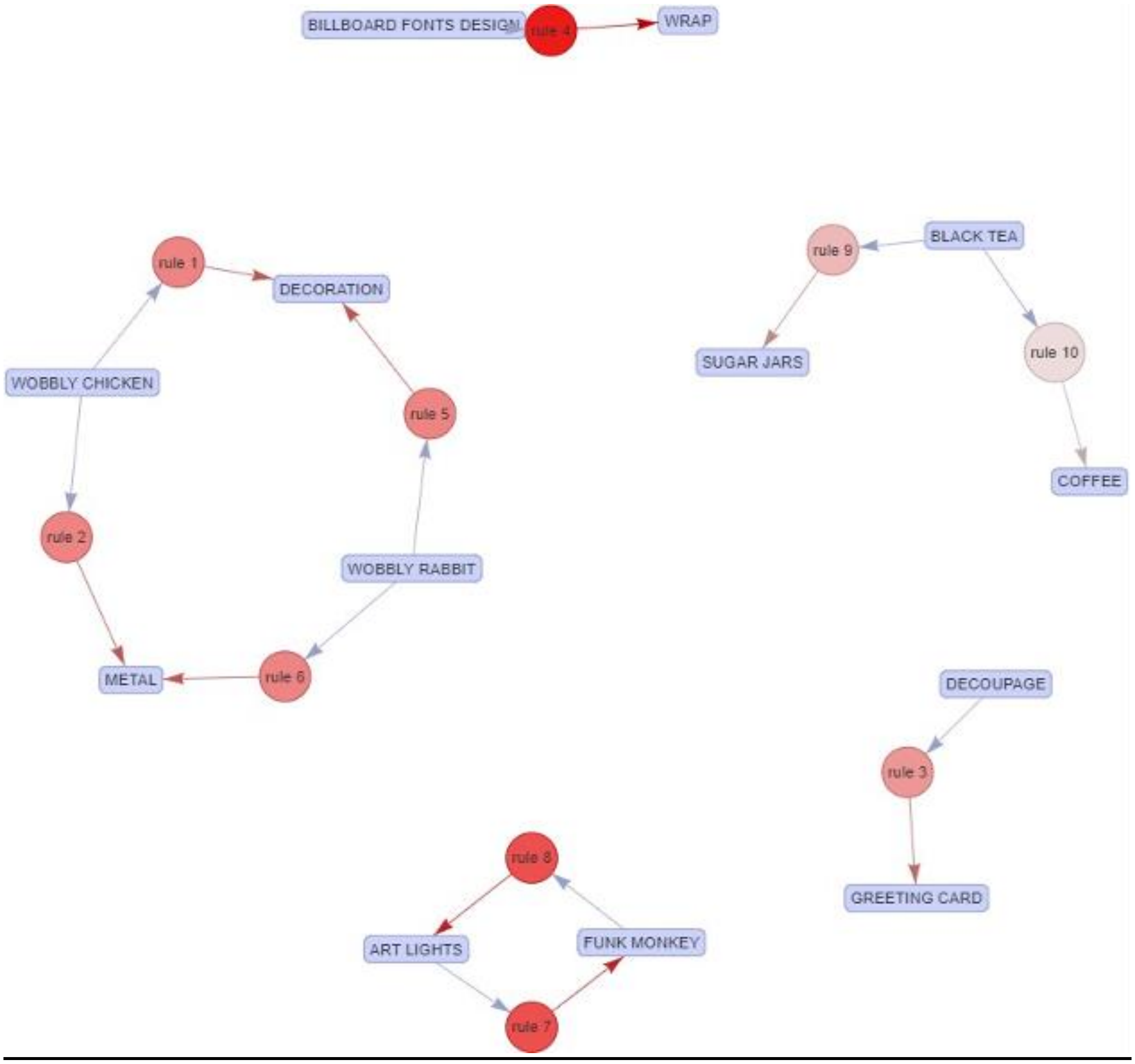

BILLBOARD FONTS DESIGN
rule 4
WRAP
rule 1
DECORATION
WOBBLY CHICKEN
rule 5
rule 2
WOBBLY RABBIT
METAL
rule 6
rule 9
BLACK TEA
SUGAR JARS
rule 10
COFFEE
DECOUPAGE
rule 3
GREETING CARD
rule 8
ART LIGHTS
FUNK MONKEY
rule 7

Figure 11

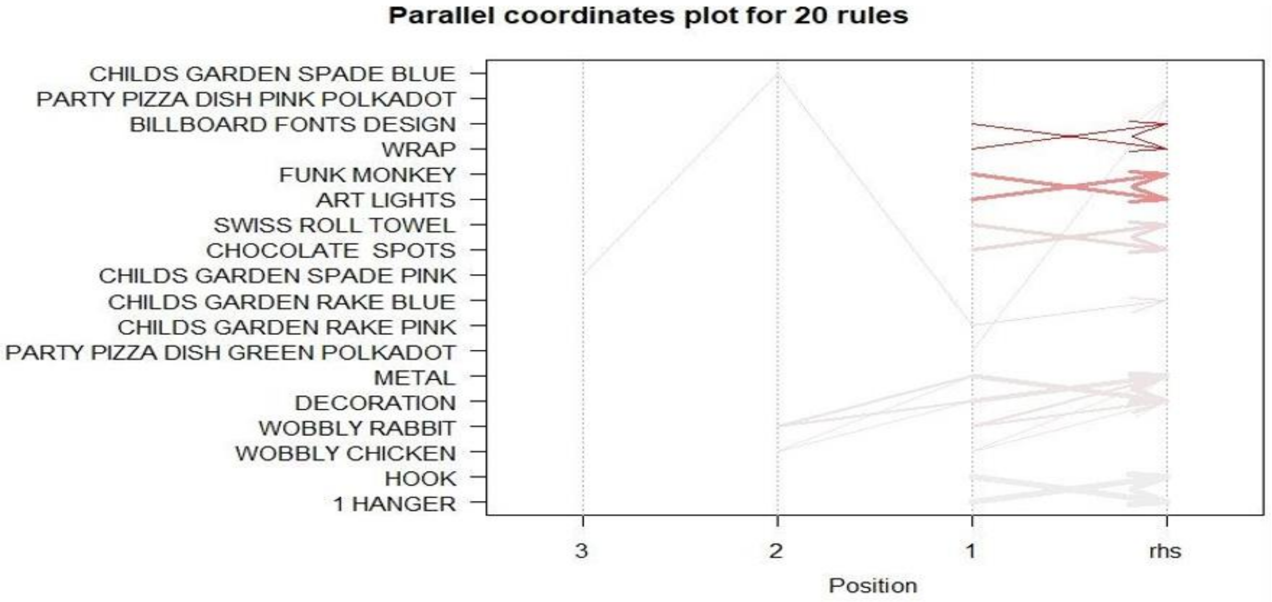


Figure 12

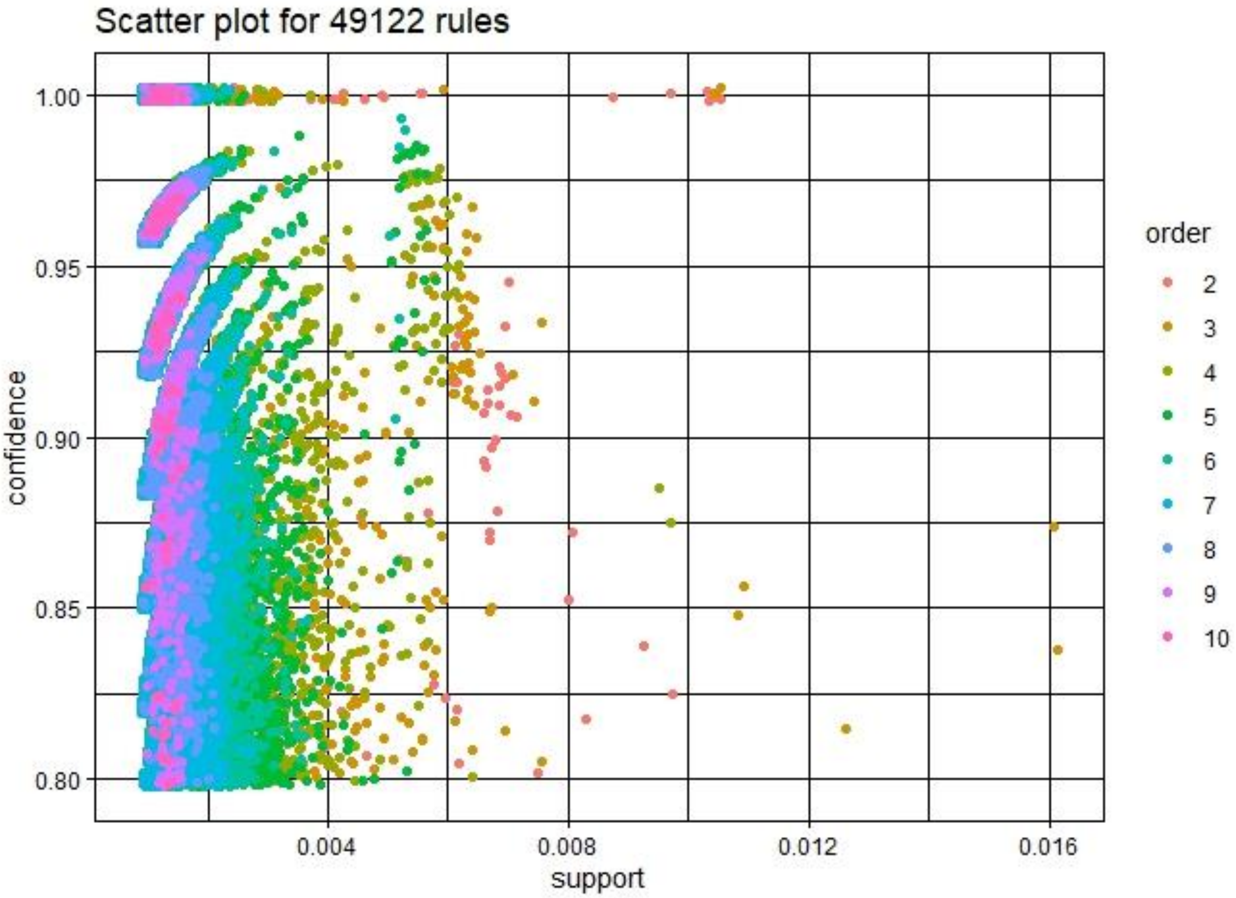
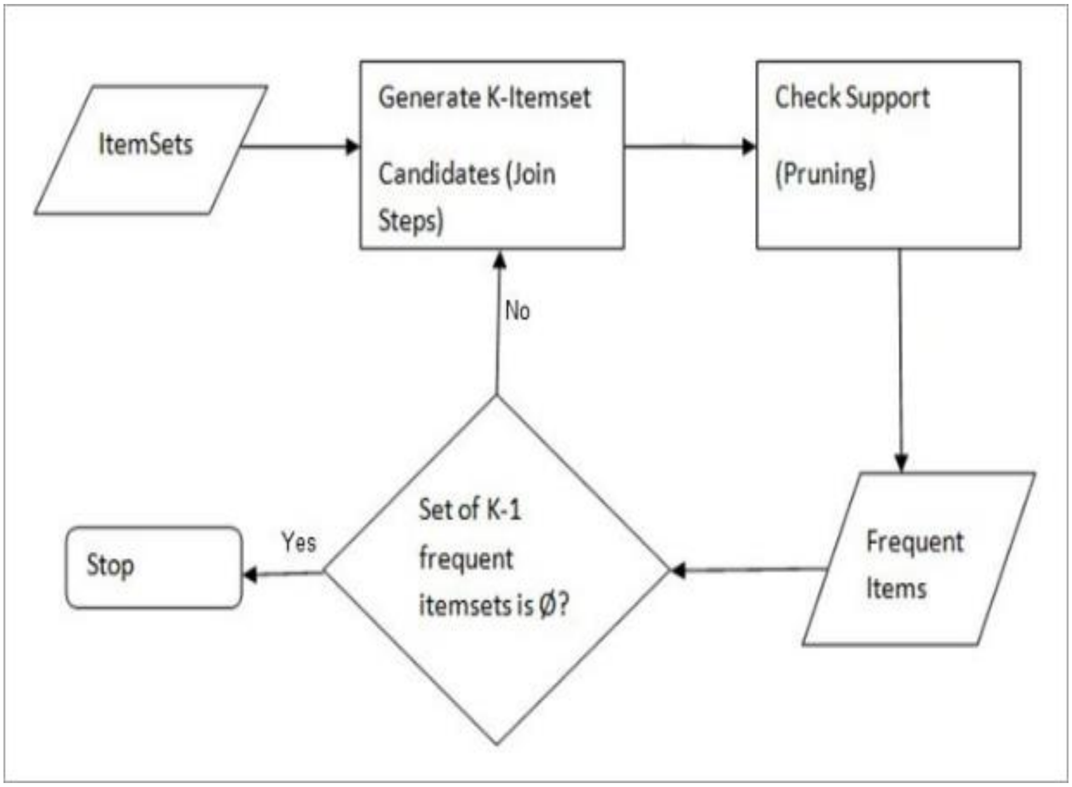


Figure 13

Chapter-5
CONCLUSION

5.1 CONCLUSION

The Apriori algorithm successfully generates descriptive information of more occurring itemsets & association rules for the data of the supermarket. The frequent data items are generated from the given input data and based on the frequent itemsets strong association rules were generated. Support and Confidence needs to be chosen wisely for the rules to be generated. Specific rules can be generated based on the requirement. In conclusion, the Apriori algorithm has generated rules effectively as per the problem statement.


ItemSets
Generate K-Itemset Candidates (Join Steps)
Check Support (Pruning)
No
Frequent Items
Set of K-1 frequent itemsets is Ø?
Yes
Stop

5.2 FUTURE SCOPE

Market Basket Analysis will help small retailers in enhancing their inventory sales by analysing and understanding the behaviour of their customers like what items they buy most together. This project can be extended to integrate in a tool or software or website where retailers can upload their excel data and can observe the customer behaviour.

REFERENCES-

[1] Agrawal, R., Fayyad, U., Kitsuregawa, M., Kotagiri, R., Kumar, V., Ooi, B., ... & Philip, S. Y. Advisory Committee DSAA 2020.

[2] Gangurde, R., Kumar, B., & Gore, S. D. (2017). Building prediction model using market basket analysis. *Int. J. Innov. Res. Comput. Commun. Eng*, *5*(2), 1302-1309.

[3] Solnet, D., Boztug, Y., & Dolnicar, S. (2016). An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. *International Journal of Hospitality Management*, *56*, 119-125.

[4] Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

[5] Kamley, S., Jaloree, S., & Thakur, R. S. (2014). An Association Rule Mining Model for Finding the Interesting Patterns in Stock Market Dataset. *International Journal of Computer Applications*, *93*(9).

[6] Kapadia, G., & Kalyandurgmath, K. (2015). Market basket analysis of consumer buying behaviour of a lifestyle store. In *International Conference on Technology and Business Management* (pp. 406-412).

[7] Al-Maolegi, M., & Arkok, B. (2014). An improved Apriori algorithm for association rules. *arXiv preprint arXiv:1403.3948*

[8] Chandwani, M. H. (2018). Market basket analysis using association rule. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*, 744-747.

[9] Gupta, S., & Mamtora, R. (2014). A survey on association rule mining in market basket analysis. *International Journal of Information and Computation Technology*, *4*(4), 409-414.