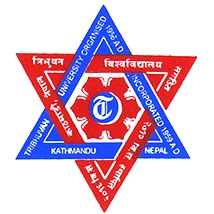
**SOCH COLLEGE OF IT**



**Tribhuvan University**

**Institute of Science and Technology**

**Pre Final Project Report**

**on**

**Market Basket Analysis:**

A desktop based application for finding “frequent bought together”

**Submitted to**

**Department of Computer Science and Information Technology**

**Soch College of IT**

*In partial fulfillment of the Requirement for the Bachelor Degree in Computer Science and Information Technology*

**Submitted by:**

**Anil Akheli**

January, 2021

ABSTRACT

Market basket analysis is also known as association rule learning, is a data mining technique that can be used to analyzing to find out if we buy an item then we are trying to buying or not-buying a group (or single) items. It is a technique which is used for discovering interesting the relationship between purchased items in retails. Market Basket Analysis system provide an information to the retail organization about the relationships and establish patterns across purchases. The purpose of this analysis is to produce a set of rules that associate two or more products together. The case of supermarkets is addressed here using frequent item-set mining. The project uses file as database. The frequent item sets are mined from database using the Apriori algorithm and then the association rules are generated. This project is gainful to the supermarket/malls, to determine the relationship between the items that are purchased by their customers.

**Keywords:** Market Basket Analysis, Association Rule Mining, Apriori Algorithm

TABLE OF CONTENTS

[CHAPTER 1: INTRODUCTION 1](#_Toc62074479)

[1.1. Background 1](#_Toc62074480)

[1.2. Problem Statement 2](#_Toc62074481)

[1.3. Objectives 2](#_Toc62074482)

[1.4. Scope and Limitations 3](#_Toc62074483)

[1.5. Report Organization 3](#_Toc62074484)

[CHAPTER 2: REQUIREMENT ANALYSIS AND FEASIBILTY ANALYSIS 5](#_Toc62074485)

[2.1. Literature Review 5](#_Toc62074486)

[2.2. Requirement Analysis 7](#_Toc62074487)

[2.3. Feasibility Analysis 8](#_Toc62074488)

[2.4. Requirement Determination and Methodology 11](#_Toc62074489)

[2.5. Data Modeling 14](#_Toc62074490)

[CHAPTER 3: SYSTEM DESIGN 16](#_Toc62074491)

[3.1. Class Diagram 16](#_Toc62074492)

[3.2. Activity Diagram 18](#_Toc62074493)

[3.3 Interface Design 19](#_Toc62074494)

[CHAPTER 4: IMPLEMENTATION AND TESTING 20](#_Toc62074496)

[4.1. Implementation 20](#_Toc62074497)

[4.2. Testing 23](#_Toc62074498)

[CHAPTER 5: CONCLUSION 25](#_Toc62074499)

[CHAPTER 6: References 26](#_Toc62074500)

**INDEX OF TABLES**

Table 1. Gantt chart…………………………………………………………………9

Table 2. Sample Data……………………………………………………………….10

Table 3. Correction of output……………………………………………………….22

**INDEX OF FIGURES**

Figure 1. Mapped into items ….…………….…...………………………………….. 11

Figure 3. Use Case Diagram……………………. …………………………………. 15

Figure 4. Class diagram ……………………………………………………………..17

Figure 5. Activity diagram ……………………………………………..……………19

Figure 6. Interface Design…………………………………………………………... 20

# INTRODUCTION

## Background

Market basket analysis applies association rules learning to purchase data to extract association or co-occurrence from a store’s transactional data and identifying cross-selling opportunities. For example, when a customer buys an item in a supermarket all the details about their purchase history go into the transaction database. Later, this huge data of customers are analyzed to determine the purchasing pattern of customers. Also, decisions like which item to stock more, cross-selling, up-selling, shelf arrangement are done.

Association rule mining (ARM) identifies the association or relationship between a large set of data items and forms the base for market basket analysis. Association rule mining has been widely used in various industries besides supermarkets, such as mail order, telemarketing production, fraud detection of credit-card, e-commerce. Association Rules are widely used to analyses retail basket or transaction data and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.

One of the major challenges for companies that have invested heavily in customer data collection is how to extract vital information from their vast customer database and product feature database to gain a competitive advantage. Market basket analysis has been largely used in many companies as a means to discover product associations.

Market basket analysis is one possible way to find out which items can be put together. It helps to identify the purchasing behavior of the customer. By mining, the data from the large transaction database shop managers can study the behavior, pattern or buying habits of the customer to increase the sale. In Market Basket Analysis, you look to see if there are combinations of products that frequently co-occur in a transaction.

In the retail industry, Market basket analysis explores the relationship between products by considering the co-occurrence of purchases in previous transactions. Item association does not necessarily suggest a cause and effect, but simply a measure of co-occurrence.

## Problem Statement

There are many Malls and Supermarkets in cities to makes customers easy to buys goods and needs. Customers are going to Malls/Supermarkets to search and buy different items. The problem many retailers face is the placement of the items. They are unaware of the purchasing habits of the customer. So they don’t know which items should be placed together in the store. With the help of this application shop managers can determine the strong relationships between the items, which ultimately helps them to put products that co-occur together close to one another. Also decisions like which item to stock more, cross-selling, up-selling, store shelf arrangement are determined.

## Objectives

The large number of transaction dataset are wasted in Nepalese markets. But they are very crucial for increasing the trading sale of retails/market. I will be creating software named Market basket analysis a desktop based application for finding frequent brought together.

### General objective

The general objective of Market Basket Analysis (MBA) system are as follows

* To analysis large datasets of purchase history of a customer, to reveal product grouping, as well as products that are likely to be purchased together by.

### Specific Objectives

* To Increase Return on investment of cross-sell targeting programs.
* To identify the next product that might interest a customer.
* To generate the association property on items from the frequent item sets.
* To identify the next product that might interest a customer.

## Scope and Limitations

### Project Scope

The scope of MBA is a limited to desktop-based application whose main aim is to simplify and improve the efficiency of the managing the goods and items. This application is targeted towards a supermarkets of Nepal.

### Limitation of the System

* This system does not recommend the other product that are not in the market. It only generate the relation between the items that are sold.
* By the analysis of history of items transactions it only tells/show about items relation and it show the customer behavior of purchasing.

## Report Organization

This study has been organized into following five chapters:

In chapter 1, I introduce why my system is built i.e. the Background, objectives of the project and its features along with its scope and limitation.

In chapter 2, I review the existing literature that contains reviews of journals and articles, and earlier thesis related to this subject. I also discuss about the related algorithms i.e, apriori algorithms along with Requirement determination as well as feasibility analysis of the system. The data modeling and process modeling technique are used to give the information about the system Requirement.

In chapter 3, I discuss about the system design. The system design can be class diagram, activity diagram, data modeling, interface design and process design

In chapter 4, I discuss about the development methodology and which tools are used on my project to make it possible. The testing is also explained in this part.

In chapter 5, I discuss about the conclusion of the project.

# REQUIREMENT ANALYSIS AND FEASIBILTY ANALYSIS

## Literature Review

Data Mining is a process used by companies to turn raw data into useful information which provides a lot of opportunities in the market sector. By using software to look for patterns in large number of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales, decrease costs. Understanding the behavior of the customer has become a vital and challenging problem for the retailers to support in this competitive environment. The challenge that the organization faces is to extract the information from their vast customer databases, to gain a competitive advantage.

Yanthy [1]et al in their paper author states about the important goal in data mining are to reveal hidden knowledge from data and various algorithms have been proposed, but the problem is that typically not all rules are interesting an only small fraction of the generated rules would be of interest to any given users. Hence numerous methods such as confidence, support, and lift have been proposed to determine the best or most interesting rules. However, some algorithms are good at generating rules high in one measure but bad in another.

Rakesh Agarwal [2] proposed the Apriori algorithm. Apriori was the first associative algorithm proposed and future development in an association, classification, associative, classification algorithms have used apriori as part of the technique. Apriori algorithm is a level-wise, breadth-first algorithm that counts transactions. Apriori algorithm uses previous knowledge of frequent item-set properties. Apriori uses an iterative approach known as a level-wise search, in which n itemsets are used to explore (n+1) itemsets. To improve the efficiency of the level-wise generation of frequent item-sets Apriori property is used here. Frequent itemsets are those items whose support value is greater than the threshold value or user-defined minimum support value.it means if A and B are the frequent itemsets together, then individually A and B should also be the frequent itemsets.

Apriori property insists that all non-empty subsets of a frequent item-sets must also be frequent.

It is one of the Data Mining Algorithm which is used to find the frequent items set from a given data repository. The algorithm involves 2 steps:

a. Pruning

b. Joining

The Apriori property is the important factor to be consider before proceeding with the algorithm Apriori property states that If an item A is joined with item B then,

Support (A=>B) = P(A U B)

Basically when I am determining the strength of an association rule i.e. how strong the relationship is between the transaction of the items, I measure through the use of the support and confidence.

The support of an item is the number of transaction containing the item. Those items that do not meet the minimum support are excluded from the further processing. Support determines how often a rule is applicable to a given data set.

Confidence is defined as the conditional probability that a transaction containing the A will also contain the B. it tells us how frequently items A and B are purchase together, when no. of times A is purchases

Confidence(A => B) = P(B|A) = P(A U B)/ P(A)

Lift indicates the strength of rule over the randomness of A and B being purchase together. It basically measures the strength of any association rule

Lift(A => B) = Confidence(A =>B)/P(B) = P(A U B)/P(A)P(B)

Phani Prasad J, Murlidher Mourya [3] in this paper author states that there are lots of case studies about the Association Rules and existing data mining algorithms usage for market basket analysis but focuses on Apriori algorithm and concludes that the algorithm can be modified and it can be extended in the future work which also decrease the time complexity. Author also clearly states the De-merits of the algorithm but claims that there is the way to improve the efficiency of the algorithm.

**Demerits**

a. It scans the database lot of times i.e. every time it runs it scans database every time, this results in shortage of memory to store the data.

b. The I/O load is not sufficient therefore it takes a lot of time to process and exhibits low efficiency.

c. The time complexity is very high.

**Solution to improve efficiency**

a. Group items into higher conceptual group e.g. white and brown bread

became “bread”.

b. Reduce the number of scan of the entire database.

In Nepalese supermarket/malls, there is not such type of specific system neither was built in markets/malls for analysis of transaction datasets. But to analysis transaction data which they stored in computer they do manually if they wanted to. And they are now starting to used desktop to keep the history and doing bill payment on digital form.

## Requirement Analysis

### Functional Requirements

The functional Requirements of this desktop based application include:

* **Data input:**

Data exploration was done manually. Data had to mapped to lists values to give input to the system.

* **Data Processing:**

A two pronged approach was taken to executing this Requirement. Frequent item set generation and association rule generation based on frequent item sets.

* **Data output:**

Report can be available in the console.

### Non-Functional Requirements

The non-functional Requirements of this desktop based application include:

* **System Availability:** The system shall be available more than 99% of the time.
* **Maintainability:** The system shall be easy to maintain.
* **Responsiveness:** The system shall respond in a timely fashion to user’s request.
* **Performance:** The system shall respond in a timely fashion and will not consume inordinate amounts of system resources.

## Feasibility Analysis

The feasibility of the system has been studied from the various aspects like whether the system is feasible technically, operationally and economically. The present technology is found to be sufficient to meet the Requirements of the system. This system is believed to work well when it is developed and installed. Hence, operational feasibility is achieved. Since the Requirements for the project are easily available, were headed with the intention to use the available resources to fulfill the system Requirement. The detail feasibility study I conducted is mentioned below.

### Technical Feasibility

The technology needed for the proposed system that I am going to develop is available. I can work for the project is done with current equipment existing tools like Tkinter, python. I can develop my system still using this technology if needed to upgrade. In future, if I want to use new technology like android app of my system it is possible. Hence, the system that I am going to develop will successfully satisfy the needs of the system for technical feasibility.

### Operational Feasibility

The user interface will be user friendly and no training will be required to use this application. The solution proposed for my project is operationally workable and most likely convenient to solve the irrelevant document and fraud e-mail

### Economic Feasibility

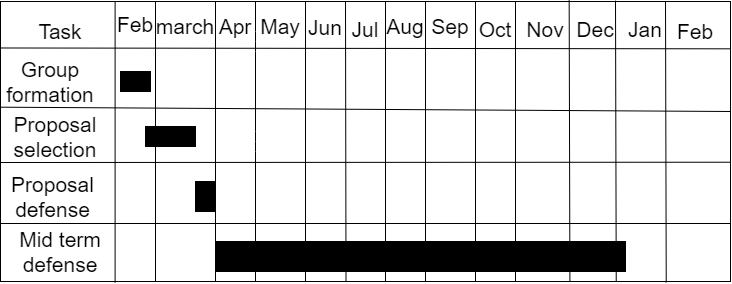
Since the system will be develop as a part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it gives an indication that the system is economically possible for development.

### Legal Feasibility

This project is developed which does not violate any legal rules and regulations. The project does not break any legal rules. So, the project is legally feasible.

### Schedule Feasibility

This is the Gantt-Chart representation of my Project Schedule on when I execute what part of my project.



##### Gantt chart

## Requirement Determination and Methodology

### Requirement collection methods

To make a real desktop based application that fulfilled the needs of Nepalese supermarkets I did different procedure to collect the required information. I went through various articles and went through different resources. I enquire to department store and shopping center and analysis the response from the retailers. For doing enquires I went to the siddhibinayak shopping center. I informally asked them how they analysis the transaction datasets, In response I found that they didn’t do any kind of analysis. I conduct more surveys like questionnaire and interviewing ahead and collect the required information about the system feature. I did questionnaire, to collect the required information about the Requirement of system development.

### Data collection

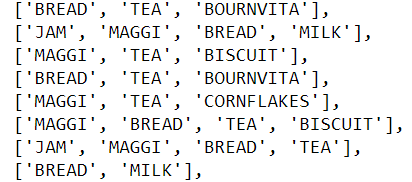
The data was collected from [http://www.salemmarafi.com.wpcontent/uploads/2014/03/ grocies.csv](http://www.salemmarafi.com.wpcontent/uploads/2014/03/%20grocies.csv) due to the unavailability of data from the supermarkets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Transaction no |  |  | items |  |
| 1. | Milk | Bread | Biscuit | Cornflakes |
| 2. | Maggi | Tea | Biscuit |  |
| 3. | Jam | Milk | Bread | Milk |

##### Sample data

### Data processing

The data collected was mapped as list as shown in figure 2. Data processing is done by importing pandas module and evaluate/read the datasets from the file. Then calculate by applying dataframe function. Using transaction encoder function, occurrence of items are assigned in the form of table labeled the value of True and False. After doing transaction encoder I convert the items in the lists and labeled. For example the collection of milk, bread and biscuit was labeled as 0 and so on.



#### Mapped into items

The listed items are then assigned by True or False, on the basis of their occurrence/available in the list. The given figure shown.

### Apriori algorithm

Association rule mining is a technique to identify underlying relations between different items. Association rules provide information of this type in the form of “IF-THEN”

Statements. The rules are computed from the data, an association rule has two numbers that express the degree of uncertainty about the rule.

1. Support(Supp)
2. Confidence(Conf)
3. Lift(L)

**Support**

The support of an item is the number of transaction containing the item. Those items that do not meet the minimum threshold support value are excluded from the further processing.

Support determines how often a rule is applicable to a given datasets.

**Support(A) = Supp(A)/total\_no\_of\_transaction**

**Confidence**

Confidence is defined as the ratio of the no of transaction that includes all items in {B} as Ill as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}

**Confidence(A=>B) = Supp(AUB)/Supp(A)**

**Lift:**

The lift of the rule A=>B is the confidence of the rule divided by the expected confidence, assuming that the item-sets A and B are independent of each other.

**Lift (A=>B) = Conf(A=>B)/Supp(B)**

**Pseudo-code**

//Find all frequent itemset

Apriori(database D of transaction, min\_support){

F1={frequent 1-itemset}

K=2

While Fk-1≠ Empty Set

Ck=AprioriGeneration (Fk-1)//Generate candidate item sets.

For each transaction in the database D {

Ct=subset (Ck, t)

For each candidate c in Ct{

Count c++

}

Fk={c in Ck such that countc&gt;min\_support}

K++

}

F=U K&gt;Fk

}

//prune the candidate item sets

Apriori generation (Fk-1) {

//Insert into Ck all combination of elements in Fk-1 obtained by self-joining item sets in Fk-1

//Delete all item sets c in Ck such that some (K-1) subset of c is not in Lk-1

}

//find all subsets of candidate contained in t

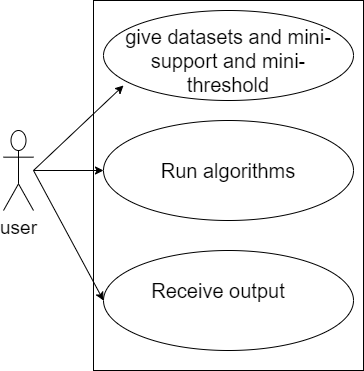
Subset (Ck, t)

}

## Data Modeling

Data modeling is a process used to define and analyze data requirements needed to support the business processes within the scope of corresponding information systems in organizations. Data modeling techniques and methodologies are used to model data in a standard, consistent, predictable manner in order to manage it as a resource. Data modeling may be performed during various types of projects and in multiple phases of projects. Data model is a detailed model that captures the overall structure of data in an organization. use case diagram are commonly used in data modeling.

### Use-Case Diagram



#### Use Case Diagram

The above mentioned diagram is the use case diagram for my project Market Basket Analysis. The user will provide input to the application .The input is a text file where items are mapped into integer’s value. The row represents the items that were purchased in one single transaction and column represents the transaction.

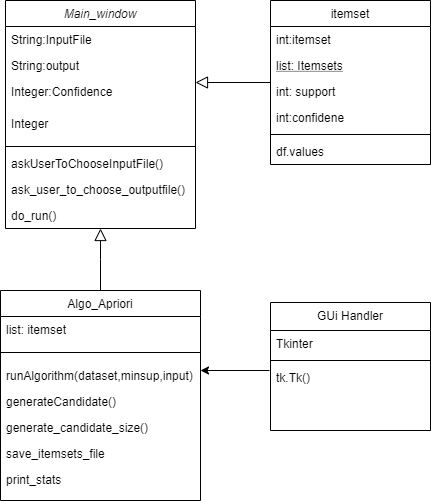
The user then can select the algorithm to run and provide the necessary parameters i.e. confidence and support.

After processing the user will receive output in to the desired path that the user wants. The output will be a text file containing association rules.

# SYSTEM DESIGN

## Class Diagram

It shows the different objects, their relationship, their behaviors, and attributes.



#### Class Diagram

The above mentioned diagram is the class diagram of Market Basket Analysis system.

The main Window class is used to present the user interface for choosing the input file and output file as desired by the user.

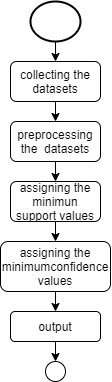
The Algo Apriori is the class that performs all the calculations once the data is provided by the user. It generates the candidate item sets and determines the size of the item sets. Finally the statistics are provide to the user in the same GUI and output is written to the desired file.

The item set class stores the items as the array of integer and provides the support of the respective item from the given input data.

## Activity Diagram

The below mentioned diagram is the activity diagram of Market Basket Analysis. The Activity can be done in following manners

A datasets can be collected or extracted from database. Preprocessing the datasets and assigning the minimum support and confidence value to the dataset.



#### Activity Diagram of Market Basket Analysis

## 3.3 Interface Design

The above mentioned diagram is the interface design of Market Basket Analysis System

## 

#### Interface Design

The main aim to make interface design were to provide GUI to the user. A user can assigned or edit the support and confidence value. The transaction datasets stored in database is extracted and selected for the analysis.

# IMPLEMENTATION AND TESTING

## Implementation

Implementation phase is one of the important phases of project development. In this phase software design is translated into source code using suitable programming language. At the time of implementing, many error and issues are occurs. However, solving the every issues each modules are implemented and runs.

My project is based on the Rapid Application Development model (RAD). I used Rapid application development while developing this system salesman sit together to define and finalize project Requirements .After clearly understood the system work closely with clients to create and improve upon working prototypes until the final product is ready. Prototype is converted into working models, collect feedback from clients to make the best system. Finally the system is tested and handover to clients.

### Rapid Application Development Process

RAD is based on prototyping and iterative development with no specific planning involved.

A prototype is a working model that is functionally equivalent to a component of the product. The Prototyping Model is one of the most popularly used Software Development Life Cycle Models (SDLC models).This model is used when the customers do not know the exact project requirements beforehand. In this model, a prototype of the end product is first developed, tested and refined as per customer feedback repeatedly till a final acceptable prototype is achieved which forms the basis for developing the final product. In this process model, the system is partially implemented before or during the analysis phase thereby giving the customers an opportunity to see the product early in the life cycle. The process starts by interviewing the customers and developing the incomplete high-level paper model. Software can radically change the entire process used to solve a problem. As a result, knowledge gained from the development process itself can feed back to the Requirements and design of the solution. Plan-driven approaches attempt to rigidly define the Requirements, the solution, and the plan to implement it, and have a process that discourage changes. RAD approaches, on the other hand, recognize that software development is a knowledge intensive process and provide flexible processes that help take advantage of knowledge gained during the project to improve or adapt the solution.

The RAD model phases are:

* **Requirements planning**   - when the team agrees on the key issues and obtains management authorization to continue
* **User design** - user interact with systems analysts and develop models and prototypes that represent all system processes, input, and outputs.
* **Construction** - focuses on program and application development task.
* **Cutover** - The final tasks including data conversion, testing, changeover to the new system.

### Tools and Languages Used

For development of the desktop based application, I use:

* **Python** :

Python is a general purpose, dynamic, [high-level](https://www.javatpoint.com/classification-of-programming-languages), and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is most popular programming language that offers the flexibility and power for programmers and data scientists to perform data analysis and apply machine learning algorithms.  In recent years, Python has become more popular for data mining due to the rise in the number of data analysis libraries. I picked the most commonly used python libraries for data analysis such as Matplotlib, numpy, pandas and so on.

* **Tkinter:**

Tkinter is the actual way in Python to create Graphical User interfaces (GUIs) and is included in all standard Python Distributions. In fact, it’s the only framework built into the Python standard library. This Python framework provides an interface to the Tk toolkit and works as a thin object-oriented layer on top of Tk. The Tk toolkit is a cross-platform collection of “graphical control elements”, aka widgets, for building application interfaces. Tk widgets can be used to construct buttons, menus, data fields, etc. in a Python application. Once created, these graphical elements can be associated with or interact with features, functionality, methods, data or even other widgets.

**Other Tools**

* Text-Editor (Pycharm/ VS Code)

I used text editor pycharm and vs code for writing the code and develop the system. PyCharm is an [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) (IDE) used in [computer programming](https://en.wikipedia.org/wiki/Computer_programming), specifically for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) language. pycharm have many feature such as Coding assistance and analysis, with code completion, syntax and error highlighting, linter integration, and quick fixes, Integrated Python debugger. Similarly vs code is a free source code editor, feature include support for debugging, syntax highlighting intelligent code completion.

* Jupyter notebook:

The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. Major feature of the jupyter notebook is the ability to display plots that are the output of running code cells. Jupyter notebook combine two component:

* Web application:

**A web application** a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.

* Notebook documents:

A representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects

## Testing

Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provide a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner.

### Testing correctness of the output

In this test approach sample input was given to the system and different support and confidence was provided. At first large sample input were given to the system with different support and confidence value. Then few sample input were given to the system with different support and confidence.

**Test case 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.**  **No.** | **Test Inputs** | **Expected Output** | **Actual Output** | **Result** |
| 1. | Item-sets = GroceryStoreDataSet.csv  Min\_support = 0.15  Min\_condifence =0.6 | Displayed the item-sets of having more than min\_supp and min\_conf. | 17 | Test  Successful |
| 2. | Item-sets = GroceryStoreDataSet.csv  Min\_support = 0.2  Min\_condifence =0.6 | Displayed the item-sets of having more than min\_suppport and min\_confidence. | 4 | Test Successful |
| 3. | Item-sets =  Min\_support = 0.1  Min\_condifence =0.8 | Dataset is not found | Select some dataset. | Test  Successful |

##### Correctness of output

# CONCLUSION

Market Basket Analysis aims to find relationships and establish patterns across purchases. The purpose of this analysis is to generate a set of rules that link two or more products together. Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository. This system is very useful for the retailers. They can use this application for the improvement in their selling strategy,

# References

|  |  |
| --- | --- |
| [1] | T. S. K. Y. W. Yanthy, Mining Interesting Rules by association and Classification Algorithms FCST 09. |
| [2] | ”. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large," *Journal of Computer Science and Technology,* vol. 15. |
| [3] | M. M. Phai Prasad J, "Market Basket Analysis Using a Data Mining," *International Journal of Emerging Technology and Advanced Engineering,* vol. 3, no. 6, june 2013. |