MARKET BASKET ANALYSIS BASED ON FREQUENT ITEM SET MINING

PROJECT REPORT

Submitted by

KARTHIK V NAIR (TVE16IE037)

AFEEF T V (LTVE16IE065)

SHAZIN ASHRAF (TVE16IE056)

BONJOE TOM ISAIAH (TVE16IE026)

to

the APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the Degree

of

Bachelor of Technology

in

Industrial Engineering

Department of Mechanical Engineering

College of Engineering, Trivandrum

May 2020

DECLARATION

I undersigned hereby declare that the project report **Market basket analysis based on frequent item set mining**, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of **Prof Mahesh S**. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously used for the award of any degree, diploma or similar title of any other University.

Place: Trivandrum

Date: June 18, 2020

Karthik V Nair

Afeef T V

Shazin Ashraf

Bonjoe Tom Isaiah

DEPARTMENT OF MECHANICAL ENGINEERING
COLLEGE OF ENGINEERING, TRIVANDRUM

CERTIFICATE

This is to certify that the report entitled **"MARKET BASKET ANALYSIS BASED ON FREQUENT ITEM SET MINING",** submitted by **Karthik V Nair**, **Afeef T V** , **Shazin Ashraf**, **Bonjoe Tom Isaiah** to the **APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY** in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Industrial Engineering is a bonafide record of the project presented by them under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Prof. Mahesh S	**Dr. Anil Lal**
Assistant Professor	Professor and Head
Department of Mechanical Engineering	Department of Mechanical Engineering
(Guide)	(Head of Department)

ACKNOWLEDGEMENT

It is with extreme respect that we remember the names of all who had been a great help and guidance throughout our project. With a profound sense of gratitude, I would like to express my heartfelt thanks to my guide, **Prof Mahesh S**, Assistant Professor, the Department of Mechanical Engineering for his expert guidance, co-operation and immense encouragement in pursuing this seminar. I am very much thankful to **Dr.Anil Lal**, Head of the Department of Mechanical Engineering, for providing necessary facilities and his sincere co-operation. My sincere thanks is extended to all the teachers of the Department of Mechanical Engineering and to all my friends for their help and support.

Karthik V Nair

Shazin Ashraf

Afeef T V

Bonjoe Tom Isaiah

ABSTRACT

Market Basket Analysis is an important part of the retail organization's analytical system for determining the placement of products, designing sales strategy for specific consumer groups to increase customer loyalty and therefore the supermarket income. This project was done using association rule mining on the transactional data of a retail supermarket in Kazhakootam,Trivandrum.Factors such as support,lift,confidence were used to determine the relationship between the products.This project is beneficial for supermarket managers to determine the relationship betweenthe items that are purchased by their customers.A Graphical User Interface(GUI) desktop application was also developed for the store as part of the project.

Contents

List of Figures

List of Tables

ABBREVIATIONS

1. ARM-ASSOCIATION RULE MINING
2. GUI-GRAPHICAL USER INTERFACE
3. MBA-MARKET BASKET ANALYSIS

Chapter 1

Introduction

1.1 Background

Market Basket Analysis is a method of data mining that focuses on discovering consumer purchasing trends by extracting associations or co-occurrences from transactional data from a shop. For example, when the individual checkout products in a supermarket ,they go into the transaction database with all the information about their purchase. Later, this huge amount of customer data is analyzed to determine customer buying patterns. Also decisions are made such as which item to stock more, cross sale, up sale, store shelf arrangement. Market basket analysis has been intensively used in many companies as a means to discover product associations and base a retailer's promotion strategy on them.Informed decision can be made easily about product placement, pricing, promotion, profitability and also find out, if there are any successful products that have no significant related elements. Similar products can be found so those can be placed near each other or it can be cross-sold.

A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together. Market basket analyses gives retailer good information about related sales on group of goods on the basis customers who buys bread often also buy several products related to bread like milk, butter or jam. It makes sense that these groups are placed side by side in a retail center so that customers can access them quickly. Such related groups of goods also must be located side-by-side in order to remind customers of related items and to lead them through the center in a logical manner.

Association rule mining (ARM) describes the association or relationship between a broad set of data objects, and forms the basis for analysis of market baskets. Aside from supermarkets, association rule mining is widely used in various industries, such as mail order, telemarketing production, credit card fraud detection and e-commerce. How to extract important information from their vast customer databases and product feature databases to gain competitive advantage is one of the challenges for companies that have invested heavily in customer data collection. In many companies market basket analysis has been intensively used as a means of discovering associations of products.Market basket analysis determines the products which are bought together and to reorganize the supermarket layout, and also to design promotional campaigns such that products purchase can be improved. Hence, the Market consumers behavior needs to be analyzed, which can be done through different data mining techniques.Data mining finds interesting patterns from databases such as association rules,correlations and many more, of which the mining of association rules is one of the most popular problems. Association rule mining finds interesting association or correlation relationships among a large set of data items.A retailer must learn and respond to customer needs. Study of market baskets is one potential way of figuring out which products to bring together. By mining the data from the huge transaction database,shop managers can study the customer's behavior or buying habits to boost the sales.

1.2 PROBLEM STATEMENT

People today buy daily goods from nearby supermarkets. There are many supermarkets which supply their customers with goods. Many retailers face the problem of putting the items in the shop. They are unaware of the customer's buying habits so they don't know which items to put together in their store. Using this application shop managers can determine the strong relationships between the items that ultimately helps them to put together products that co-occur close to each other. Decisions such as which item to stock more, cross sell, up sell, store shelf arrangement are also determined.

1.3 OBJECTIVE

- To identify the frequent item set from the transaction on the basis of factors such as support , confidence,lift
- To find relationship between the products and determine the placement of products in supermarket
- Analyse customer buying patterns and determine promotion strategy.

1.4 SCOPE

The project was done using the transactional data of BUY N SAVE supermarket in Kazhakootam,Trivandrum.T data collected was pre-processed before analysing the relationship between the products.Products were categorised and association rule mining was done in python environment.And a desktop application was also developed for the supermarket for the ease of use based on python GUI.

1.5 LIMITATIONS

- The project was done on the basis of 1500 transactions in the supermarket ,more transactional data could result in better relations among the products.
- Due to limited amount of data,many similar products of various brands were categorised.If more transactional data was obtained the association among different product brands could have been more clear.
- The purchasing pattern of the customers can vary due to a lot of factors which include seasonality,price and demography which can affect the obtained result.

Chapter 2

Literature Review

Association rule mining finds interesting association or relationships among a large data set. Association rules are derived from the frequent itemsets using support and confidence as threshold levels. The sets of items which have minimum support are known as Frequent Itemset. The support of an itemset is defined as the proportion of transactions in the data set which contain the itemset. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern. Association rules are derived depending on the confidence. Frequent itemset generation is done mostly using data mining Apriori algorithm .

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant [1] and came to be known as Apriori. This algorithm uses two steps “join” and “prune” to reduce the search space. It is an iterative approach to discover the most frequent itemsets. Apriori algorithm is a level-wise, breadth-first algorithm which counts transactions. Apriori algorithm uses prior knowledge of frequent item set properties. Apriori uses an iterative approach known as a level-wise search, in which n-item sets are used to explore (n+1) - item sets. To improve the efficiency of the level-wise generation of frequent item sets Apriori property is used here. Apriori property insists that all non-empty subsets of a frequent item set must also be frequent. This is made possible because of the anti-monotone property of support measure - the support for an item set never exceeds the support for its subsets. The Apriori property is the important factor to be consider before proceeding with the algorithm. Apriori property states that If an item X is joined with item Y,

$$Support(X \cup Y) = min(Support(X), Support(Y))$$

Basically when we are determining the strength of an association rule i.e. how string the relationship is between the transaction of the items we 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 an item A will also contain the item B. Confidence(A –> B)

$$P(A/B) = P(A \cap B)/P(A) = support(A \cup B)/support(A).$$

In some cases confidence can give misinformation on the dependency between the items. Because in confidence ,the popularity of the consequent item is not taken into consideration. A third metric called lift solves this problem. Lift illustrates how likely item B is purchased when item A is purchased, while controlling for how popular item B is.

- If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
- If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
- If the lift is < 1, that lets us know the items are substitute to each other. This means that presence of one item has negative effect on presence of other item and vice versa.

2.1 STEPS IN APRIORI

- In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

- Let there be some minimum support, min-sup . The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min sup are taken ahead for the next iteration and the others are pruned

- Next, 2-itemset which are frequent items with min-sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

- The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only

- The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min-sup. If all 2-itemset subsets are frequent then the superset will be frequent, otherwise it is pruned.

- Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min-sup criteria. The algorithm is stopped when the most frequent itemset is achieved.

2.2 EXAMPLE OF APRIORI

Let the database of transactions consist of following itemsets as show in Table2.1 (a). We will use Apriori to determine the frequent item sets of this database. To do this, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value 3 is the support threshold. The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately. By scanning the database for the first time, we obtain the result as shown in Table 2.1(b). All the itemsets of size 1 have a support of at least 3, so they are all frequent. The next step is to generate a list of all pairs of the frequent items. For example, regarding the pair 1,2: the first table of Example 2 shows items 1 and 2 appearing together in three of the itemsets; therefore, we say item 1,2 has support of three as shown in

Table 2.1(c). The pairs 1,2, 2,3, 2,4, and 3,4 all meet or exceed the minimum support of 3, so they are frequent. The pairs 1,3 and 1,4 are not. Now, because 1,3 and 1,4 are not frequent, any larger set which contains 1,3 or 1,4 cannot be frequent. In this way, we can prune sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs. In the example, there are no frequent triplets. 2,3,4 is below the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.

Item
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

(a)

ITEM	SUPPORT
{1}	3
{2}	6
{3}	4
{4}	5

(b)

ITEM	SUPPORT
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	5

(c)

ITEM	SUPPORT
{2,3,4}	2

(d)

Table 2.1: Example of Apriori

Chapter 3

System Design

3.1 METHODOLOGY

3.1.1 Data collection

The data was collected from BUY'N'SAVE marginfree bazaar in MENAMKULAM , KAZHAKUTTAM THIRUVANANTHAPURAM on 06/02/2020 for the days 01/02/2020 TO 05/02/2020.

BUY'N'SAVE marginfree bazaar

MENAMKULAM, KAZHAKUTTAM THIRUVANANTHAPURAM

E-Mail: buynsavebazaar@gmail.com

BILL WISE SALES STATEMENT FROM 01/02/2020 TO 05/02/2020

BILL NO.	AR	TY NAME			BILL VALUE
0059230	CASH	()			40.00
	1	8902102126270	FAST VASH L S	2	16.60 20.00
0059231	CREDIT	CARD ()			242.00
	1	8906094750338	MURLYA MILK P	1	58.00 58.00
	2	8906014080774	BRHMNS PEPR A	2	39.55 41.00
	3	8908008820243	CRUNCHY KARAS	1	41.90 45.00
	4	100055	APPLE NEV	330	180.00 180.00
0059232	CREDIT	CARD ()			657.00
	1	8906087470724	2QR ASHOK REG	1	115.00 145.00
	2	8902102125495	HENKO V/P 3KG	1	365.00 440.00
	3	8901088050586	PAR OIL 175ML	1	65.00 63.00
	4	VONDER	VONDER HOT GL	2	7.50 10.00
0059233	CASH	()			30.00
	1	89005637	NESTLE KITKAT	2	12.59 15.00
0059234	CASH	()			138.00
	1	M	BLUE MILMA	6	23.00 23.00
0059235	CREDIT	CARD ()			91.00
	1	8906009991061	ELITE FAMILY	1	45.00 45.00
	2	M	BLUE MILMA	2	23.00 23.00

Figure 3.1: Data Sample

3.1.2 Data pre-processing

The data collected was then sorted out to get a valid amount of frequency in each category according to the products properties.

EXAMPLES OF CATEGORIES

ICECREAM	STATIONARY
AMULVANILA M	HAUSER RUSH G
VESTA CHCO BA	HAUSER BILLI
VESTA FRUIT F	YATHRA
VESTA JUBO KU	ZIGZAG BLK (M
MERRIBOY MILK	KALIKUDUKKA
AMULTRI CONE	MINNAMINNI
MERRIBOY FAMI	SMALL WONDER
MERRIBOY CONE	MALAYALAMANOR
MERRIBOY TUB	PREMIER HANKY
MERRIBOY MANG	PENFIELD NOTE
MERRIBOY SUND	CAR MAT GODRA
AMULMANGO DU	STEELSCALE D
AMULKOOLFI M	THOZHILVARTHA
AMULSTRWBRY	MAGICPOT
MERRIBOY PIST	HAUSERGERMNY
	HAUSERACTIVE
	FOSCALES 15C
	MANORAMA SAMB
	VANITHA
	FEVOCOLSQUEE
	CLSMATE N BK
	CHART PAPER
	SAFTEY PIN
	CLARUS POCKET
	CM BOCTAINE
	CM GEL PENOC
	CM RULED NOTE

Figure 3.2: Example of Categories

3.1.3 Apriori algorithm

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attribute value conditions that occur frequently together in a given dataset. A typical and widely used example of association rule mining is Market Basket Analysis. For example, data are collected from the supermarkets. Such market basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. 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.

a. Support

b. Confidence

Support 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 further processing. Support determines how often a rule is applicable to a given data set.

$$Support(X \cup Y) = min(Support(X), Support(Y))$$

Confidence Confidence is defined as the conditional probability that a transaction containing the LHS will also contain the RHS. Confidence (LHS->RHS)

$$P(RHS/LHS) = P(RHS \cap LHS)/P(LHS) = support(RHS \cap LHS)/support(LHS)$$

Confidence determines how frequently item in RHS appears in the transaction that contain LHS. While determining the rules, we must measure these two components as it is very important to us. A rule that has very low support may occur simply by chance.

Lift is a ratio of observed support to that expected if X and Y were independent. In other words, lift illustrates how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

- If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

- If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

- If the lift is < 1, that lets us know the items are substitute to each other. This means that presence of one item has negative effect on presence of other item and vice versa.

Therefore having lift bigger than 1 is critical for proving associations

$$Lift(X => Y) = supp(X \cup Y)/(supp(X) * supp(Y))$$

3.2 System Design

3.2.1 Class diagram

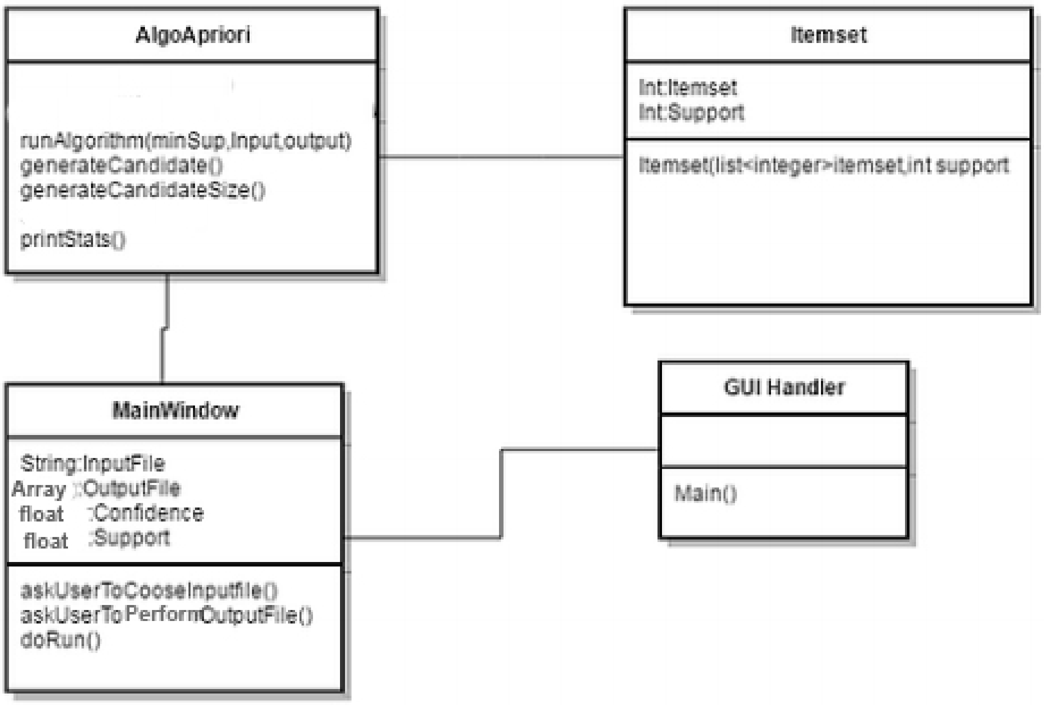

Figure 3.3: Class diagram

As shown in Figure 3.3, there are three main classes used in the application The MainWindow class is used to present the user interface for choosing the input file and output file as desired by the user. The AlgoApriori 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 and also finds the support, confidence and lift. 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.

3.2.2 Sequence diagram

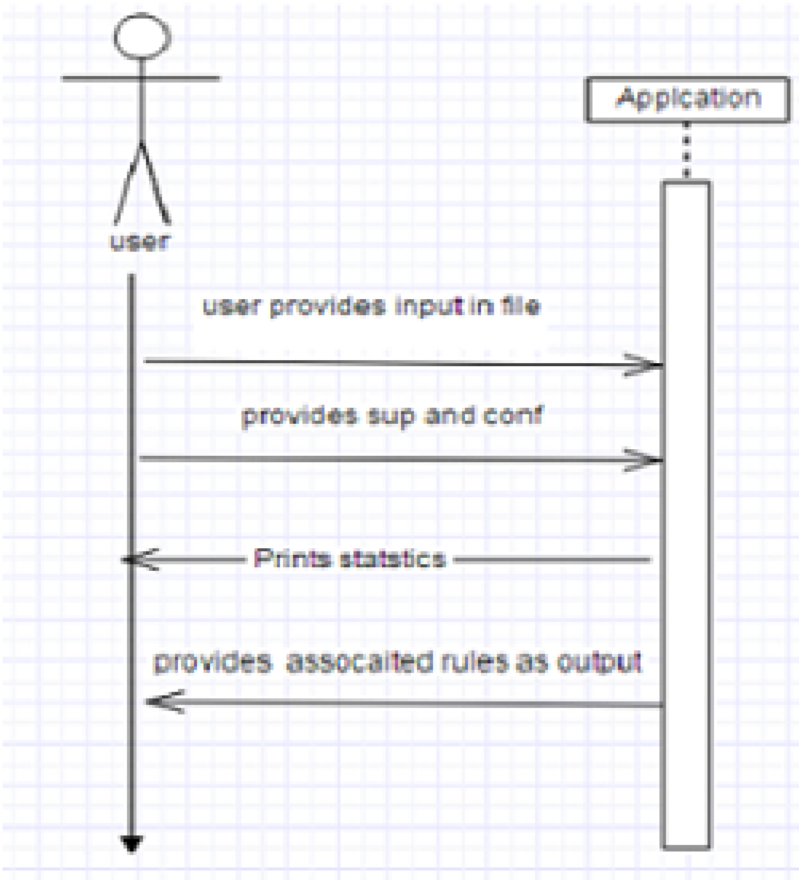

Figure 3.4: Sequence diagram

3.3 IMPLEMENTATION

3.3.1 Extracting the most frequent itemsets

Our first step in the implementation of the algorithm is to extract the most frequent itemsets by giving a threshold value of 0.01 for the support measure. This is done with the help of mlxtend library.

```
frequent_itemsets = apriori(dataset, min_support=0.01, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x: len(x))
frequent_itemsets
```

Figure 3.5: Extracting the most frequent itemsets

3.3.2 Extracting the rules which are dependent to each other

We can create our rules by defining metric and its threshold. In this step we will set our metric as “lift” to define whether antecedents and consequents are dependent or not. Threshold is selected as 1.1 since it is required to have lift scores above than 1 if there is dependency.

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.1)
rules["antecedents_length"] = rules["antecedents"].apply(lambda x: len(x))
rules["consequents_length"] = rules["consequents"].apply(lambda x: len(x))
rules.sort_values("lift",ascending=False)
```

Figure 3.6: Extracting the rules which are dependent to each other

3.3.3 Finding strong association rules

This is the most important step in which our aim is to find the strong association rules. We use confident metric to extract the strong association rules from the whole transactions.

```
rules.sort_values("confidence",ascending=False)
```

Figure 3.7: Extracting the rules which are dependent to each other

3.3.4 Finding other interesting relationships

In order to find the itemsets which are strongly associated with carry bag, we extract the association rules in which carry bags appearing as consequent.

```
strong_rules=rules[rules["consequents"].str.contains("bag", regex=False)& rules["consequents_length"]==1 ]
strong_rules.sort_values("confidence", ascending=False)
```

Figure 3.8: Finding other interesting relationships

Chapter 4

GUI Application

A graphical user interface(GUI) desktop application using python was developed for the BUY N SAVE supermarket for easiness of the supermarket managers .

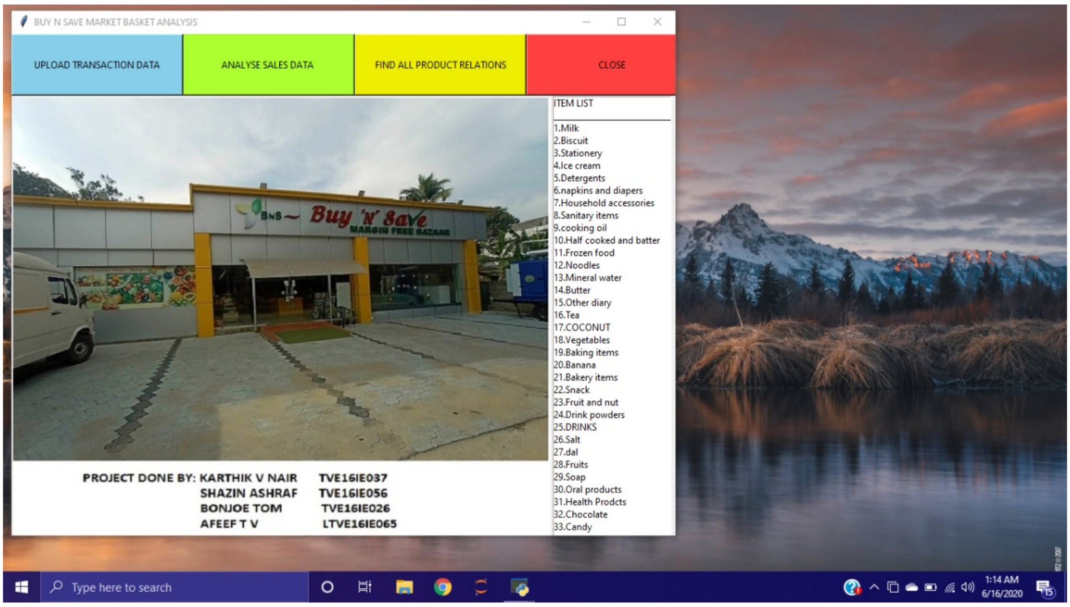

Figure 4.1: Application screenshot

Different functions available in the application is listed below.

4.1 UPLOAD FUNCTION

On clicking the"UPLOAD TRANSACTION DATA" button , a dialog box which ask the user to select the directory of the transaction data file pop up .

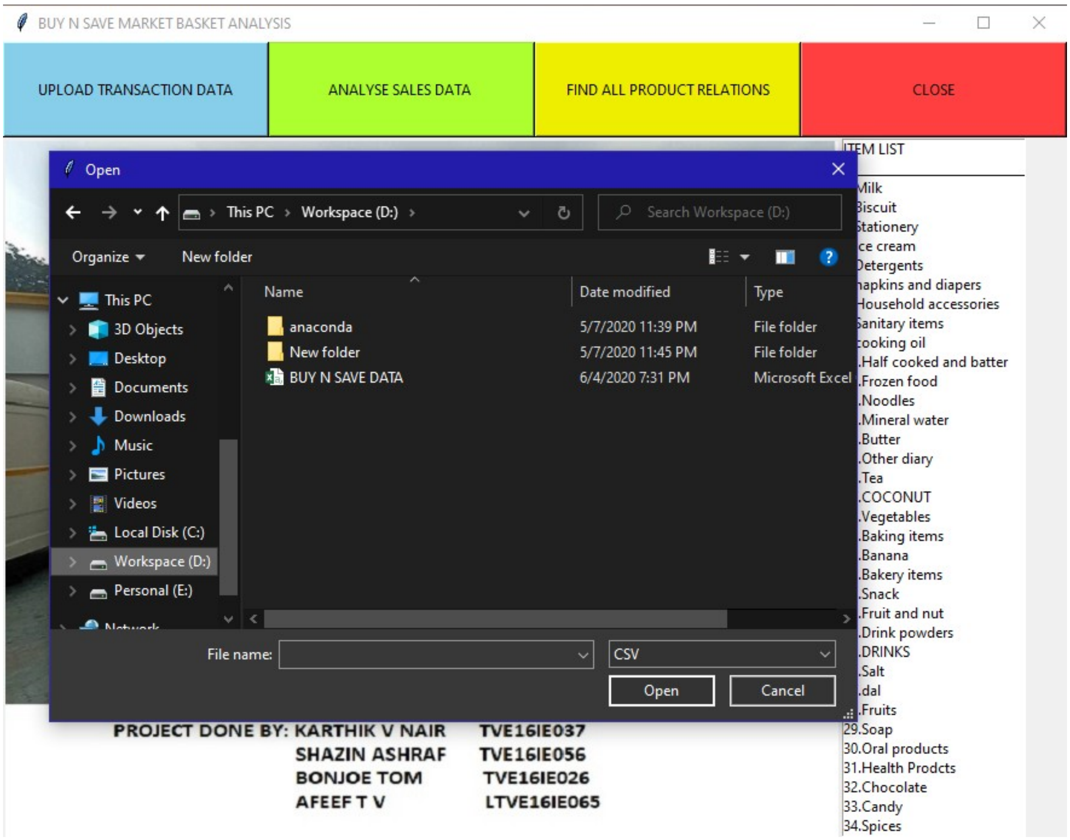

Figure 4.2: Upload function

4.2 Analyse function

On clicking the"ANALYSE SALES DATA" button , a window showing the histogram and tree diagram of the sales data opens up .

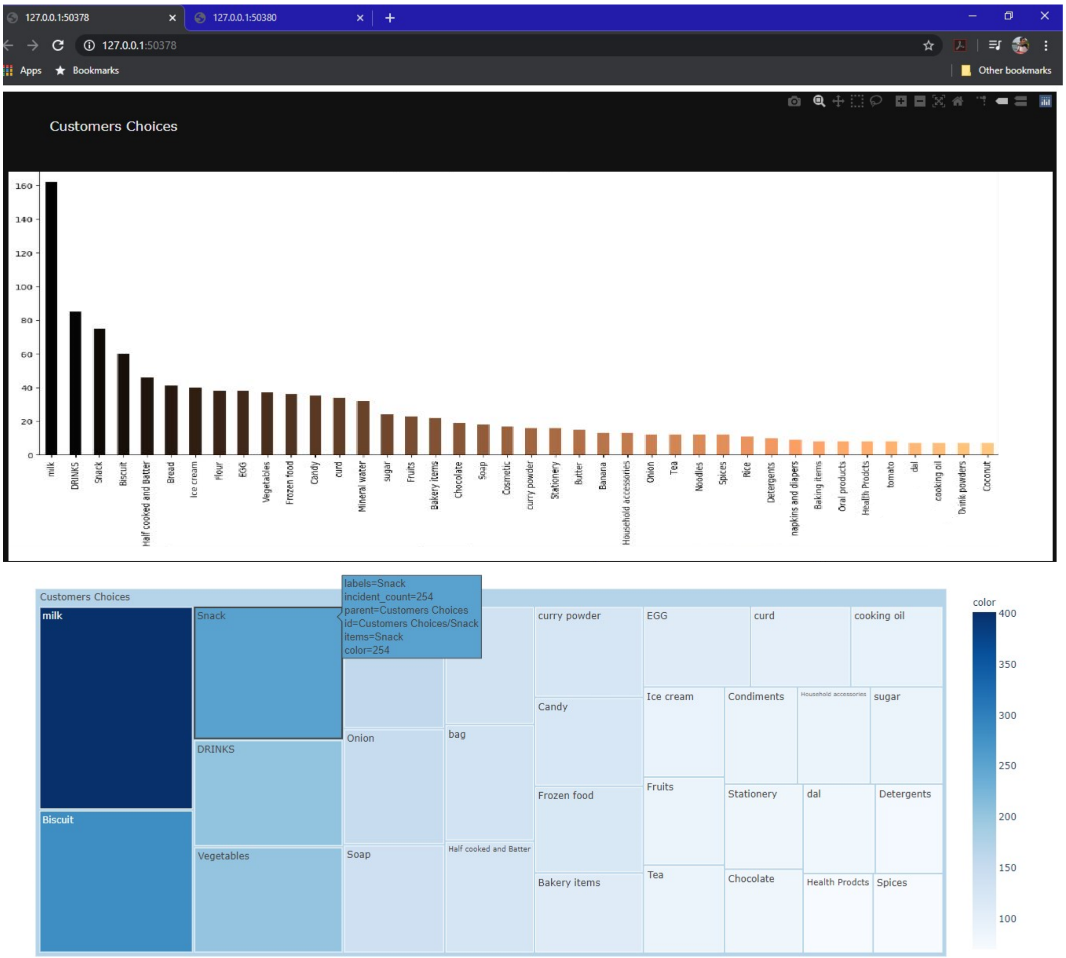

Figure 4.3: Analyse function

4.3 FIND RELATIONS FUNCTION

On clicking the"FIND ALL PRODUCT RELATIONS" button , a dialog box asking the user to choose the directory for saving the relationships table appears and the file is saved in chosen location.The user can access the file containing relations from the chosen location.

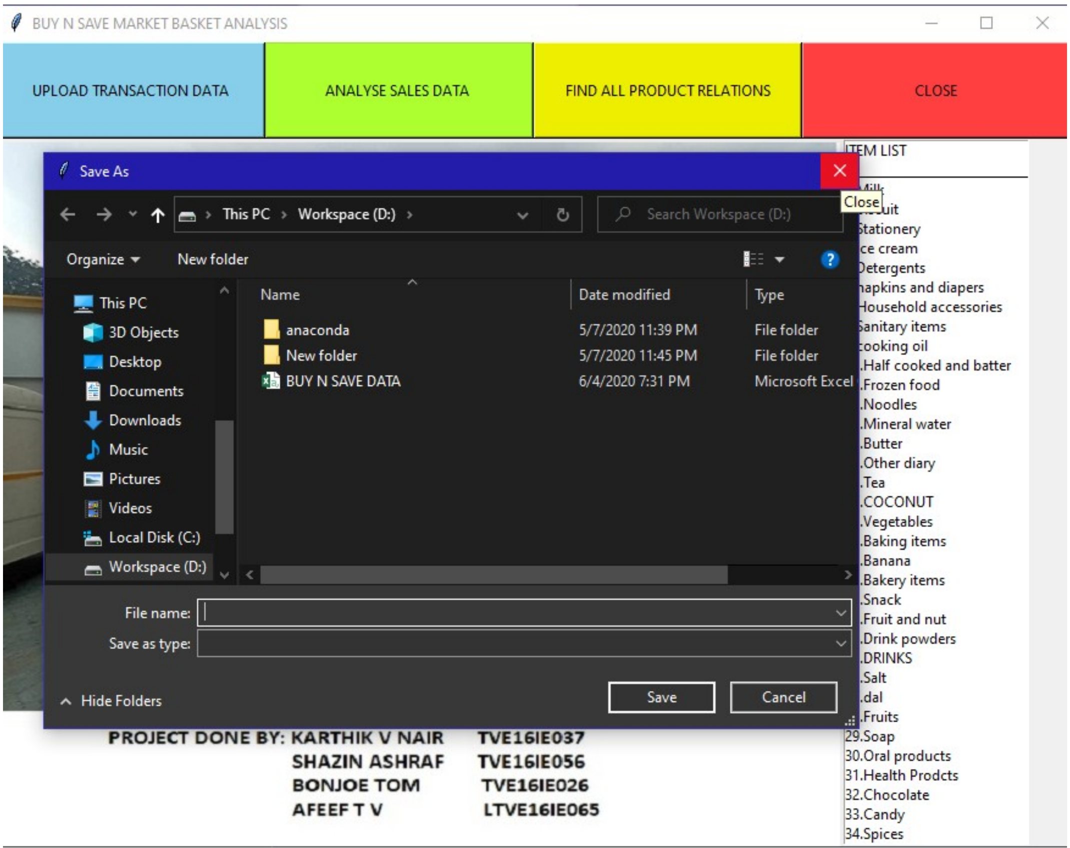

Figure 4.4: Find Relations Function

4.4 Quit Function

On clicking the"CLOSE" button , a dialog box asking the user confirmation to quit the application appears,If" Yes "is clicked application gets closed automatically ,If "No" is clicked the application returns to the normal screen.

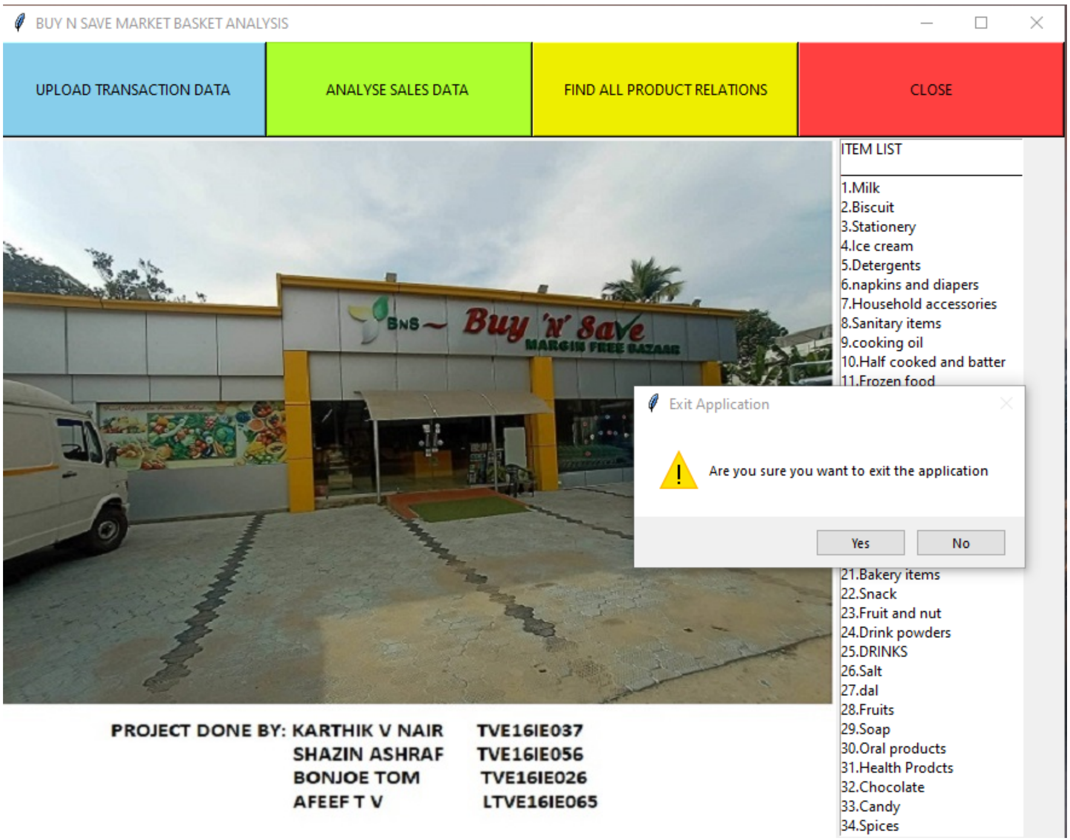

Figure 4.5: Quit Function

Chapter 5

Results and Discussion

5.1 Most Frequent Items

The most popular items are milk, drinks, snacks, biscuit, half cooked foods, batters, bread and egg etc. This gives us an idea about the buying behaviour of the customers at buy n save super market which is mainly dominated by the IT professionals staying nearer to the supermarket. So the store can focus more on these items and give exciting offers for attracting these customers.

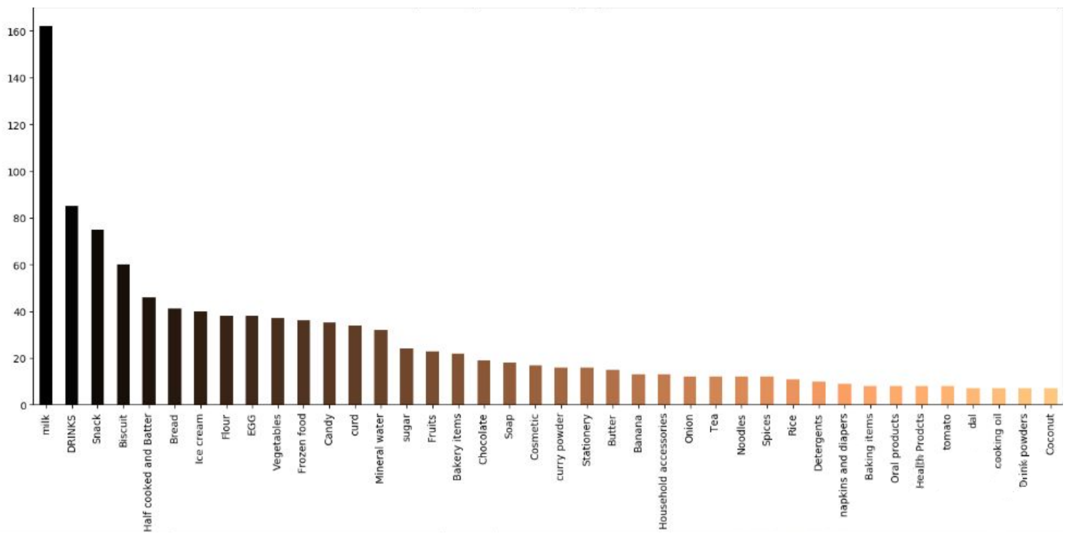

Figure 5.1: Frequency of most popular items

5.2 SOME STRONG RELATIONSHIPS

For finding the item sets with strong relationships we are focusing mainly on the confidence values. The table below shows the rules with higher confidence values. The table is obtained after ensuring a support of atleast 1 percent, the lift is greater than 1 and finally a confidence value greater than 40 percent.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
22	(cooking oil, Condiments)	(curry powder)	0.015374	0.067785	0.010482	0.681818	10.058575	0.009440	2.929819
20	(curry powder, cooking oil)	(Condiments)	0.017470	0.060098	0.010482	0.600000	9.983721	0.009432	2.349755
21	(curry powder, Condiments)	(cooking oil)	0.017470	0.058001	0.010482	0.600000	10.344578	0.009469	2.354997
0	(Biscuit)	(Biscuit)	0.022362	0.156534	0.013277	0.593750	3.793108	0.009777	2.076224
14	(Vegetables, cooking oil)	(Biscuit)	0.018169	0.156534	0.010482	0.576923	3.685611	0.007638	1.993647
17	(Flour, Onion)	(Biscuit)	0.020964	0.156534	0.011880	0.566667	3.620089	0.008598	1.946460
7	(Biscuit, DRINKS)	(Snack)	0.024458	0.145353	0.013277	0.542857	3.734753	0.009722	1.869541
12	(curry powder, Vegetables)	(Biscuit)	0.019567	0.156534	0.010482	0.535714	3.422353	0.007419	1.816696
19	(Onion, bag)	(Biscuit)	0.020266	0.156534	0.010482	0.517241	3.304341	0.007310	1.747180
16	(Biscuit, Onion)	(Flour)	0.024458	0.093641	0.011880	0.485714	5.186994	0.009589	1.762365
1	(POTATO)	(Vegetables)	0.024458	0.115304	0.011880	0.485714	4.212468	0.009060	1.720242
15	(Biscuit, Flour)	(Onion)	0.025157	0.085255	0.011880	0.472222	5.538934	0.009735	1.733201
13	(Biscuit, cooking oil)	(Vegetables)	0.022362	0.115304	0.010482	0.468750	4.065341	0.007904	1.665310
11	(Biscuit, curry powder)	(Vegetables)	0.022362	0.115304	0.010482	0.468750	4.065341	0.007904	1.665310
3	(Biscuit, Vegetables)	(milk)	0.036338	0.250175	0.016771	0.461538	1.844865	0.007681	1.392533
2	(POTATO)	(Onion)	0.024458	0.085255	0.011181	0.457143	5.362061	0.009096	1.685056
6	(Vegetables, Onion)	(milk)	0.023061	0.250175	0.010482	0.454545	1.816912	0.004713	1.374680
5	(Vegetables, Snack)	(milk)	0.025157	0.250175	0.011181	0.444444	1.776536	0.004887	1.349686
9	(Vegetables, Snack)	(Biscuit)	0.025157	0.156534	0.011181	0.444444	2.839286	0.007243	1.518239
8	(DRINKS, Snack)	(Biscuit)	0.030049	0.156534	0.013277	0.441860	2.822778	0.008574	1.511210

Figure 5.2: Association rules

23 strong association rules were obtained. One strong association rule looks like this :

- Cooking oil, condiments => curry powder (support: 6.7 percentage, confidence: 68.2 percentage)

This rule means 6.7 percent of all the transaction contain these three items together and it is evident that cooking oil, condiments, and curry powder is having a very strong relationship. Almost 70 percent of the people who purchased cooking oil and condiments also bought curry powder. If we set lower minimum support and confidence, more association rules might be generated. Also the run time will be a little longer. In other words, raising the support threshold and confidence will have a secondary effect of reducing computation time, which may be desirable for large data sets.

5.3 SOME OTHER INTERESTING RELATIONSHIPS

- Drinks, biscuit => snacks (support: 1.3 percent, confidence: 54 percent)

- Flour, onion =>biscuit (support:1.5 percent, confidence:56 percent).

  From our analysis we noticed that carry bags appears in a good number of purchases. So we decided to investigate on the items which can be the primary reason to buy these carry bags. So we try to extract the strong association rules in which carry bags appearing as a consequent. The table below shows top 10 associations with carry bag.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedents_length	consequents_length
471	(Biscuit, Onion)	(bag)	0.024458	0.083159	0.010482	0.428571	5.153661	0.008448	1.604472	2	1
344	(cooking oil)	(bag)	0.058001	0.083159	0.015374	0.265060	3.187405	0.010551	1.247505	1	1
446	(Biscuit, Snack)	(bag)	0.041230	0.083159	0.010482	0.254237	3.057257	0.007054	1.229401	2	1
348	(Tea)	(bag)	0.055206	0.083159	0.013277	0.240506	2.892139	0.008687	1.207174	1	1
270	(Onion)	(bag)	0.085255	0.083159	0.020266	0.237705	2.858452	0.013176	1.202738	1	1
336	(Frozen food)	(bag)	0.073375	0.083159	0.016073	0.219048	2.634094	0.009971	1.174004	1	1
356	(Detergents)	(bag)	0.048917	0.083159	0.010482	0.214286	2.576831	0.006414	1.166889	1	1
346	(Fruits)	(bag)	0.059399	0.083159	0.012579	0.211765	2.546515	0.007639	1.163157	1	1
334	(curry powder)	(bag)	0.067785	0.083159	0.013976	0.206186	2.479425	0.008339	1.154982	1	1
350	(Household accessories)	(bag)	0.053809	0.083159	0.010482	0.194805	2.342573	0.006008	1.138658	1	1

Figure 5.3: Some other interesting relationships

The itemset biscuit and onion is the one which makes strong relationship with the carry bag. 42 percent of the people who purchased biscuit and onion also bought carry bag.

Above are only a few relationships which we have founded. By continuous investigations we can find more interesting patterns which will be highly useful for the supermarket to market their products.

Chapter 6

Conclusion and Future Scope

The Apriori algorithm effectively generates highly informative frequent itemsets and association rules for the data of the supermarket. The frequent data items are generated from the given input data and based on the frequent item stets strong association rules were generated.

Application can be efficiently used by using more efficient algorithm rather that Apriori Algorithm in future.Integrating this software in the supermarket ERP system can yield real-time interpretations which could be used to tailor offers according to the market demand .From the result the demand of products can be forecasted to increase the efficiency of the inventory management.

References

[1] Fast Algorithms for Mining Association Rules in Large Databases *R. Agrawal and R. Srikant," Fast Algorithms for Mining Association Rules in Large Databases", Journal of Computer Science and Technology, vol. 15*

[2] Mining Interesting Rules by association and Classification Algorithms *W. Yanthy, T. Sekiya, K. Yamaguchi,, FCST 09.*

[3] Market Basket Analysis *RoshanSharma ,https://www.kaggle.com/roshansharma/market-basket-analysis*