

## Mining of Retail Transactional Data for Revealing Seasonal Facts

**Darshna B. Rajput**

Research Scholar,

Shri Jagdish Prasad Jhabarmal Tibrewala University,  
Jhunjhunu, Rajasthan

### Abstract :

Retail market generated so many daily transactions throughout the day. This transactional data has been maintained by the bill information given to the consumers. This data set also maintains the facts regarding date and time for the particular transaction. In this paper, the goal is to mine the retail transaction data and find different facts regarding the particular season that are, WINTER, SUMMER and MONSOON. For finding these facts, the need of identifying frequent item sets is required. This frequent item sets find using classification and association techniques of data mining. The Weka tool is used for this analysis and applying different methods of classification and association. According to the research, season wise different facts regarding item groups and items in super store have been revealed which is helpful for predicating consumer behavior and product buying patterns in particular season.

**Keywords :** Retail, Data Mining, Consumer, Classification, J48, Association, FP-Growth, Season

### 1. Introduction

Real world retail transactional data set is very large data set available. It is considered as one of the biggest data set. There are so many facts hidden in that data set. To reveal all this facts, mining of this retail dataset is necessary. Mining of this retail transaction data set gives the knowledge about which products sold more or less, which products sold together, which products required advertising or promoting, store layout, stock management of different products at which time period, etc. To acquire this type of knowledge from the larger retail dataset, data mining techniques are used.

Meaning of Data mining is to find out and present useful knowledge from large amounts of data to the world [1]. It is the method in which the outside world sees the data from different angles and gets the large amount of useful results. As and when time goes, the data will be increasing very fast but the fruitful information will be decreased [1]. So, the main goal of data mining technique is to find out significant knowledge from the big data set.

The purpose of this research paper is to mine the retail transaction data and find out different buying products patterns according to the season which are

WINTER, SUMMER and MONSOON. For finding this, at first classification of the whole sales data set into three different classes according to season has been done. After this, association rules for all three classes is generated which gives the frequent item sets in all different seasons. This predicts the facts about item groups, which one is the most sold item group in particular season. It also finds out that which items, stock have to be maintained according to the season.

This paper comprises of different sections which ultimately fulfill the goal of finding seasonal facts using classification and association techniques. The sections are as follows : Section 2 is about the old researches related with this topic, Section 3 explains about classification and association techniques, Section 4 gives the outline of the dataset used in the research, actual experimentation and analysis is explain in the Section 5 and in the last section, I concludes the discussion with the outcome.

### 2. Review of Literature

There are some of researches with data mining in the area of retail industry. Each research gives the facts related with retail data that detecting the hidden

knowledge from the generated pattern. In this section, I am analyzing the exiting work that already done.

D. Vidhate (2014) [2] gives the analysis of different researchers in the development of super bazaar industry. For this development, the understanding of consumer behavior is necessary. The main goal of this research paper is by studying all national and international materials, find out new trends and techniques in the area of super market.

Dr. M. Dhanabhakyaam and Dr. M. Punithavalli (2011) [3] talks about Association Rule Mining which helpful in finding frequently buying items, individually and collectively, that improves the sales of the retail shop. This association also useful in marketing, catalog designing and decision making process. The authors have used Apriori algorithm with some modification for finding association of products.

R. Chauhan et. al. (2010) [4] gives the explanation about different data mining techniques for the purpose of promoting items via e-retailing. The predictive analysis has been done through clustering and visualization. The authors applied k-means clustering in Weka data mining tool.

P. Prasad and Dr. L. Malik (2011b) [5] used Weka tool for applying clustering technique that divided segments according to consumer profile in retail sector. This helps to find out consumer behavior and buying patterns which were used to predict the next profit in the retail market.

P. Prasad and Dr. L. Malik (2011a) [6] does association rule mining for analyzing the sales pattern of the retail store. They have used Apriori algorithm in Weka tool. These rules were helpful, in decision making process. They also developed application in Visual C#, for easy GUI interface.

Sonia and Dr. R. Gargi (2014) [7] at first, do the association on super market data to find out related products and then classify the data. They used J48 algorithm for classification. They proved that the proposed algorithm is 13% more accurate than the existing one.

M. Muthurengan and R. Hepzibah (2013) [8] explains about rating of new product in the market. They have taken Walmart transaction data for the analysis. For ranting the products, they have considered the criteria like, total sales of product, total consumers that buy particular product, increasing in product selling, association of product, etc. Product rating has been done by association rule

mining as a preprocessing task and then applying Hungarian algorithm.

H. Zeb et. al. (2011) [9] performs the questionnaire survey on female consumers of Pakistan to study the buying behavior according to product brands. The survey done by five point scale measurement. The correlation has been found by SPSS software. As a conclusion the authors found that brand affects the sale but by considering the quality, position and fashion trend.

T. Hilage and Dr. R. Kulkarni (2011) [10] find the result of different sales pattern using association rule mining, rule induction and Apriori algorithm. They have done the comparative study of these three techniques. By finding the association, the outcome was helpful to retailer for planning promotional offers and managing store layout.

A. Gupta and C. Gupta (2010) [11] gives the facts regarding profit gaining through consumer purchases using the concept of data mining. They develop the application which sends mail to the interested consumers. These classification of interested consumers for promotional offers, done by clustering techniques. The authors analyzed about different problems of consumer behavior which directly or indirectly affect the sale and production process.

K. Swamy et. al. (2015) [12] explains the working environment of Weka with Association rule mining. For association, the authors used Apriori algorithm. They have taken the online shopping data of 60 persons and by setting min support and min confidence, found the result. The result has been given the frequent item set from all the correlated items.

P. Tanna and Dr. Y. Ghodasara (2014) [13] have taken the data set from the shopping center with only 15 transactions, to find out association using Apriori algorithm and Weka tool. They achieved 20 rules which were useful for finding correlated items.

A. Bala et. al. (2016) [14] have taken two association rule mining algorithm Apriori and FP-Growth for finding the patterns of correlated items. They have applied comparison of these two algorithms for deciding which one is better, by taking the criteria execution time and scan time. The result has been given that FP-Growth is faster than Apriori, by Weka tool.

B. Sabitha et. al. (2014) [15] concentrates on the factor of time period that is, which time period

morning or evening better for shopping. For decision making, the authors have used algorithms like ID3 and C4.5 in Tangra tool. They analyzed ROC curve for the result which gives the idea of true positive and false positive rate.

K. Venkatachari (2016) [16] finds the relationship of different product of super store. By applying Apriori and FP-Growth algorithm in R and Rapid Miner tool, the author mined the data of Mumbai Retail store. The collected data was in format of transactional data set which was converted to binary format. FP-Growth is better in the prospective of speed but for larger data set it may be slower.

H. Kaur and K. Singh (2013) [17] done the research on the sports data from the retail market. They applied FP-Growth algorithm for association rule mining and finding patterns of products sold together. This helps in increasing the sale of products. Different categories of sports taken by the authors were Cricket, Tennis, Badminton and Hockey. The data set contains 1 and 0 value, that is, 1 for presence and 0 for absence of item.

### 3. Classification and Association

Classification is the method for classifying data into different classes. This method divided into two phases : Learning and Classification. In the first phase, the training data set has been taken and analysis has been done on training data set. In the second phase, testing has been performed to check the accuracy of the rules. This also predicates the future outcomes. From the collected data, the classification technique is used to classify the objects according to the class labels. The classification algorithm builds a model using training set and according to model classified the new objects [18]. There are classification algorithms like Decision Tree, Neural Network, Rule base Induction, Bayesian Network and Genetic algorithm.

Association is the method for finding rules according to the relationships between products. The rules is like “If the consume bought product x, he/she also bought product y”. This means the product y is correlated with the product x. The well known example of association rule mining is Market Basket Analysis [19]. There are two major factors in association rule mining, Support and Confidence. The Support of any transaction X is calculated as the proportion of transaction in dataset which contains item set X i.e.  $SUPP(X) = X \cup Y$ . The confidence is

calculated as the proportion of the transactions that contains X which also have Y i.e.  $CONF(X \Rightarrow Y) = SUPP(XUY) / SUPP(X)$ [20]. There are many different algorithms available for association rule mining like Apriori, FP-Growth, Tertius, SETM, etc.

### 4. Retail dataset of Super store

Retail data set has been created by the Daily Grocery shopping from the super store. This data set was generated from the billing data of the super store given to the individual consumers. So, it was secondary data that the conducted research used for the experiments. This data set contains the information like bill number, bill date, item number, item name, item group, quantity, price, etc. In this research, the taken item groups are AGARBATI, COSMETICS, BATHSOAP, DETERGENT, SHAMPOO, FOOD, TEA, etc. The major attribute for the research is BILL DATE. It contains date and time, both information of the shopping. The season for the particular transaction has been derived by this bill date information. There is month available in the bill date and according to this months, the prediction of the season has been done. There are three quarters in the 12 months. Each quarter represents particular season. MAR to JUN is of SUMMER season, JUL to OCT is of MONSOON season and NOV to FEB is of WINTER season.

The data collected is in the form of transactional data set. In which each bill number contains one or more item groups. So, there is duplication of bill number in the data set. For applying this data set, into Weka tool, the binary format of the data i.e. in form of 1 and 0 is required. For this, each different item group represented as an attribute in the data set. As a value of this attribute, 1 represents presence of item group in the bill and 0 represents absence of item group in the bill. Hence, as a preprocessing task, the conversion of transactional data set into binary tabular data set has been done. This tabular data can be easily applied in the Weka tool for classification and association.

### 5. Result and Discussion

The sample data taken for the research of super store in .ARFF file format is given in the figure 1.

No.	1: BILL_NO Nominal	2: SEASON Nominal	3: BILL_DAY Nominal	4: BILL_MONTH Nominal	5: BILL_YEAR Nominal	6: BILL_DATE Nominal	7: BILL_TIME Nominal	8: AGARBATI Nominal	9: BATHSOAP Nominal	10: COSMETICS Nominal	11: DETERGENT Nominal	12: FOOD Nominal	13: SHAMPOO Nominal	14: SOFTDRINK Nominal	15: TEA Nominal	16: TOOTHPASTEPOW Nominal
1	10	SUMMER	2	Apr	2012	02-Apr-12	09:51:51	0	0	0	0	0	0	1	0	0
2	10	SUMMER	5	May	2012	05-May-12	16:56:29	0	0	0	0	0	0	1	0	0
3	10004	SUMMER	14	May	2012	14-May-12	15:31:57	0	0	0	1	0	0	0	1	0
4	10006	SUMMER	14	May	2012	14-May-12	15:38:18	0	0	0	0	1	0	0	0	0
5	10012	SUMMER	14	May	2012	14-May-12	15:50:20	0	0	1	0	0	0	0	0	0
6	10014	SUMMER	14	May	2012	14-May-12	15:59:59	1	0	0	0	0	0	0	0	0
7	10016	SUMMER	14	May	2012	14-May-12	16:10:05	0	1	0	1	1	0	0	1	1
8	10018	SUMMER	14	May	2012	14-May-12	16:14:17	0	1	1	0	1	0	0	0	0
9	10030	SUMMER	14	May	2012	14-May-12	16:37:37	0	0	0	0	1	0	1	0	0
10	10035	SUMMER	14	May	2012	14-May-12	16:43:59	0	0	0	1	0	1	0	0	1
11	10040	SUMMER	14	May	2012	14-May-12	16:59:02	0	1	1	0	1	1	0	0	0
12	10053	SUMMER	14	May	2012	14-May-12	17:19:14	0	1	0	1	0	0	0	0	0
13	10061	SUMMER	14	May	2012	14-May-12	17:36:23	0	0	0	1	1	1	0	0	0
14	10074	SUMMER	14	May	2012	14-May-12	18:03:35	0	0	1	0	1	1	0	0	1
15	10076	SUMMER	14	May	2012	14-May-12	18:08:17	0	1	0	0	0	0	0	1	0
16	10080	SUMMER	14	May	2012	14-May-12	18:17:56	0	1	0	0	0	0	0	0	0
17	10093	SUMMER	14	May	2012	14-May-12	18:41:17	0	1	0	0	0	0	0	0	0
18	10094	SUMMER	14	May	2012	14-May-12	18:52:58	0	0	0	1	0	0	0	1	1
19	10100	SUMMER	14	May	2012	14-May-12	18:59:10	0	0	1	1	0	1	0	1	0
20	10102	SUMMER	14	May	2012	14-May-12	19:03:42	0	0	0	0	1	0	0	0	1

Figure 1 : Sample retail data set in binary format

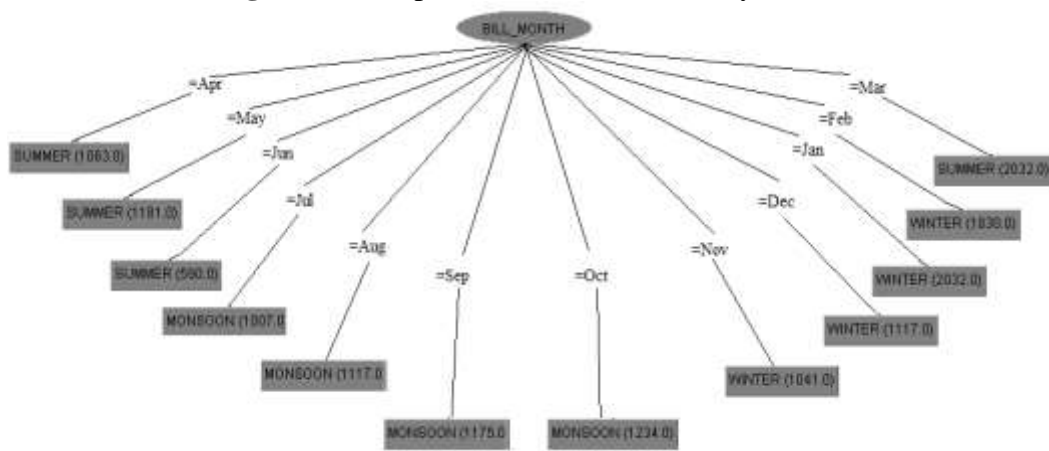


Figure 2 : Season wise classification according to BILL MONTH

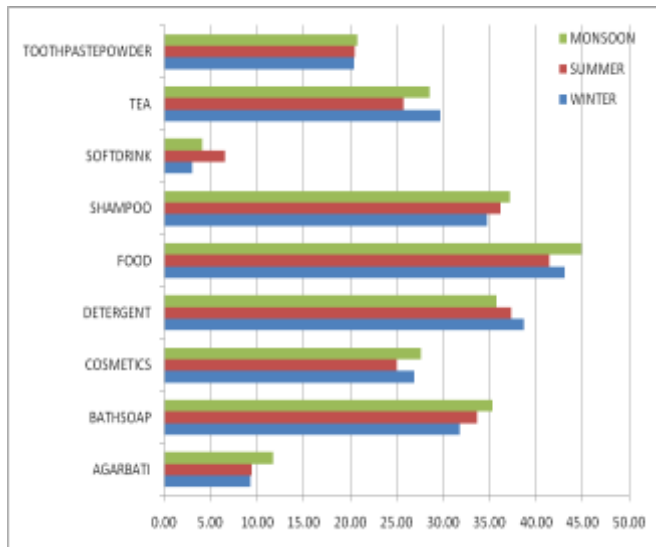
For doing seasonal analysis of this retail data, the classification of data into three different classes SUMMER, MONSOON and WINTER is required. All this analysis is useful for store manager to maintain stock of particular item group in particular season, to plan sale according to season like Monsoon Sale, Summer Bonanza, etc., to drop out stock of non-interested item groups, to provide special discount or offers, etc. All these strategies were applied by studying the past shopping data of the consumers. The collected data is of past 2 years data of the super store.

The classification of the whole data set is done according to BILL MONTH, using J48 decision tree algorithm. J48 is the cost effective option than the other classification algorithm [18]. This classification has been done with highest accuracy and record separation. The decision tree made by this classification is displayed in the figure 2.

After classifying the data, to find out frequent item group from all groups in each season, calculate support count of each group in each season. This support count value gives the maximum sold item group from the super store.

The bar chart given in the figure 3, displayed the status of each item group in different season. It clearly displays the effect of season on item group. The chart clearly said that MONSOON season has highest selling season for maximum item groups. Those groups are AGARBATI, BATHSOAP, COSMETICS, FOOD, SHAMPOO and TOOTHPASTEPOWDER. This revealed the fact that people more like to go for grocery shopping in MONSOON. TEA group has highest selling in WINTER season that presented the human urge to drink hot drink in WINTER season to be protected from cool environment. In SUMMER, people's liking toward SOFTDRINK is more. There were hot waves in SUMMER season. So, SOFTDRINK provides some relief in hot environment.

DETERGENT is more used in WINTER, than in SUMMER and at last in MONSOON. MONSOON is rainy season, in this, season the cloths take more time to be dried, so people don't like to wash their cloths more in this season. Hence, usage of detergent is low in MONSOON season.



**Figure 3 :** Seasonal effect on different item groups

After finding frequent item groups, from the retail data set, the next thing in the research is find out the pair or triplets of item groups, which were sold together. That means the next goal is to find correlated item groups from the transactional data. To find out correlation between item groups, the task performed was association rule mining. This association of different item groups has been achieved using FP-Growth algorithm. The reason behind choosing FP-Growth is its fast execution and no candidate generation [21]. This algorithm has been applied on different season data set with minimum confidence of 40% and minimum support range 1% to 100%. The association rule for WINTER, SUMMER and MONSOON are given in the table 1, 2 and 3 respectively.

**Table 1 :** Association rules for WINTER Data set

WINTER SEASON		
Sr No.	Rules	Confidence
1	DETERGENT : 2325 → FOOD : 1171	0.5
2	FOOD : 2589 → DETERGENT : 1171	0.45
3	TEA : 1785 → FOOD : 790	0.44
4	BATHSOAP : 1912 → DETERGENT : 826	0.43
5	BATHSOAP : 1912 → FOOD : 817	0.43
6	SHAMPOO : 2089 → FOOD : 878	0.42
7	SHAMPOO : 2089 → DETERGENT : 878	0.42
8	TEA : 1785 → DETERGENT : 748	0.42
9	BATHSOAP : 1912 → SHAMPOO : 765	0.4

**Table 2 :** Association rules for SUMMER Data set

SUMMER SEASON		
Sr No.	Rules	Confidence
1	DETERGENT : 1825 → FOOD : 863	0.47
2	TEA : 1257 → FOOD : 557	0.44
3	COSMETICS : 1224 → SHAMPOO : 525	0.43
4	FOOD : 2025 → DETERGENT : 863	0.43
5	TEA : 1257 → DETERGENT : 515	0.41

**Table 3** : Association rules for MONSOON Data set

MONSOON SEASON		
Sr No.	Rules	Confidence
1	DETERGENT : 1617 → FOOD : 847	0.52
2	TEA : 1293 → FOOD : 631	0.49
3	TOOTHPASTEPOW : 939 → FOOD : 455	0.48
4	COSMETICS : 1248 → SHAMPOO : 534	0.43
5	BATHSOAP : 1599 → FOOD : 672	0.42
6	SHAMPOO : 1681 → FOOD : 703	0.42
7	FOOD : 2032 → DETERGENT : 847	0.42
8	BATHSOAP : 1599 → SHAMPOO : 663	0.41
9	COSMETICS : 1248 → FOOD : 516	0.41
10	DETERGENT : 1617 → SHAMPOO : 654	0.4
11	TEA : 1293 → DETERGENT : 516	0.4
12	COSMETICS : 1248 → BATHSOAP : 498	0.4
13	BATHSOAP : 1599 → DETERGENT : 634	0.4

The above given association rules revealed the pairs of item groups, from which different items sold together. Rules analysis has given following facts :

(1) Very first rule defined the pair of item groups DETERGENT and FOOD in each season, this two group of items not directly correlated but still those items purchase together. This fact was hidden till now, so shopkeeper can organized the store layout in which the items of this two groups place at nearby places.

(2) One another rule was about TEA and FOOD group, this tells that the people like to take food items like biscuits, Gaychhap Besan, Ravo, Maida in complement with TEA. This all items are directly taken or in the form of cooking items like Bajiyas, Dhosa, Upma, Uttapam, Cake, etc. So, this rule is at almost after first position with higher confidence.

(3) One unrevealed rule is also about TEA and DETERGENT group of items. This two group items also sold together. Hence, shop keeper planned to put all three groups TEA, FOOD and DETERGENT close to each other. This easily attracts people to buy any of the group items which are unplanned with the planned group item.

(4) In MONSOON and WINTER, the selling of FOOD group items was higher, which revealed the fact of hungriness more in the season of MONSOON and WINTER.

(5) According to data, demand of TEA was higher in the season of WINTER. This gives the idea of human habitual behavior in the particular season.

(6) One more fact which was hidden is groups like DETERGENT, BATHSOAP and SHAMPOO were more required in WINTER. The selling of those group items was higher in WINTER in comparison with other seasons. This displayed the opposite behavior of the mass who shop from the super store. This fact helpful to shopkeeper for maintaining stock of those group items in WINTER season.

According to the analysis on the past retail data, using different classification and association algorithms, some hidden facts also revealed which helpful to the shop keeper for the future benefits. Therefore, using data mining techniques, the data has been mined and visualized hidden knowledge from that to the outside world.

## 6. Conclusion

In this research paper, two major algorithms J48 for making decision tree and FP-Growth for finding association rules have been used to mine the larger retail data set. Both this algorithms generated hidden facts regarding the retail data set, which were very much fruitful for the store keeper to design the store layout according to the correlated and most frequent item groups. The store keeper also planned the schemes, promotional offers or sale according to the analysis. The less sold items can be sold more by these offers. All this analysis has been helpful for increasing the sale of super store and generating more profit.

## References

1. B. Agarwal, S. Tayal, "Data Mining and Data Warehousing", University Science Press, Delhi, ISBN : 978-81-318-0658-6, 2012.
2. D. Vidate, "A Conceptual study of Consumer Behavior Analysis in Super Bazar using Knowledge Mining", Sinhgad Institute of Management and Computer Application, pp. 70-75, ISBN : 978-81-927230-0-6, 2014.
3. Dr. M. Dhanabhakya, Dr. M. Punithavalli, "A Survey on Data Mining Algorithm for Market Basket Analysis", Global Journal of Computer Science and Technology, XI (11), pp. 23-28, ISSN : 0975-4350, 2011.
4. R.[Roma] Chauhan, R.[Ritu] Chauhan, A. Goel, "Applying Data Visualization and Mining Techniques to Improve Luxury Goods Promotion through E-Retail", International Journal of Computer Applications, III (10), pp.38-42, ISSN : 0975-8887, 2010.
5. P. Prasad, Dr. L. Malik, "Generating Customer Profiles for Retail Stores Using Clustering Techniques", International Journal on Computer Science and Engineering, III (6), pp. 2506-2510, ISSN : 0975-3397, 2011b.
6. P. Prasad, Dr. L. Malik, "Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions", International Journal on Computer Science and Engineering, III (5), pp. 2177-2182 , ISSN : 0975-3397, 2011a.
7. Sonia, Dr. R. Gargi, "Apriori Based Classification using WEKA", International Journal of Engineering Sciences Paradigms and Researches, XVI (1), pp.10-14 , ISSN : 2319-6564, 2014.
8. M. Muthurengan, R. Hepzibah, "Mathematical Analysis of Retail Product Rating Using Data Mining Techniques", International Journal of Computational Engineering Research, pp. 6 -9, ISSN : 2250-3005, 2013.
9. H. Zeb, K. Rashid, M. Javeed, "Influence of Brands on Female Consumer's Buying Behavior in Pakistan", International Journal of Trade, Economics and Finance, II (3), pp. 225-231, ISSN : 2010-023X, 2011.
10. T. Hilage, Dr. R. Kulkarni, "Applications of data mining techniques to a selected business organization with special reference to buying behavior", International Journal of Database Management Systems, III (4), pp. 169-181, ISSN : (Online) 0975-5705 (Print) 0975-5985, 2011.
11. A. Gupta, C. Gupta, "Analyzing Customer Behavior using Data Mining Techniques : Optimizing Relationships with Customer", Management Insight The Journal of Inclusive Analyzers, VI (1), ISSN : 0973-936X, 2010.
12. K. Swamy, G. Babu, R. Venkatasubhai, "Identification of Frequent Item Search Patterns Using Apriori Algorithm and Weka Tool", International Journal of Innovative Technology and Research, III (5), pp. 2401-2403, ISSN : 2320-5547, 2015.
13. P. Tanna, Dr. Y. Ghodasara, "Using Apriori with Weka for Frequent Pattern Mining", International Journal of Engineering Trends and Technology, XII (3), pp. 127-131, ISSN : 2231-5381, 2014.
14. A. Bala, M. Shuaibu, Z. Lawal, R. Zakari, "Performance Analysis of Apriori and FP-Growth Algorithms (Association Rule Mining)", International Journal of Computer Technology and Applications, VII (2), pp. 279-293, ISSN : 2229-6093, 2016.
15. B. Sabitha, N. Bhuvaneshwari, G. Annapoorani, P. Balasubramanian, "Implementation of Data Mining Techniques to implement Market Analysis", International Journal of Innovative Research in Computer and Communication Engineering, II (11), pp. 7003-7008, ISSN (Print) : 2320-9798, ISSN (Online) : 2320-9801, 2014.
16. K. Venkatachari, "Market Basket Analysis Using FP-Growth and Apriori Algorithm : A Case Study of Mumbai Retail Store", BVIMSR'S Journal of Management and Research, VIII (1), pp. 56-63, ISSN : 09760-4739, 2016.
17. H. Kaur, K. Singh, "Market Basket Analysis of Sports Store Using Association Rules", International Journal of Recent Trends in Electrical and Electronics Engineering, III (1), pp. 81-85, ISSN : 2231-6612, 2013.
18. T. Patil, S. Sherekar, "Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science And Applications, VI (2), pp. 256-261, ISSN : 0974-1011, 2013.
19. J. Arora, N. Bhalla, S. Rao, "A Review on Association Rule Mining Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, I (5), pp. 1246-1251, ISSN : (Online) 2320-9801 (Print) 2320-9798, 2013.

20.N. Maheshwari, N. Pandey, P. Agrawal, “Market Basket Analysis Using Association Rule Learning”, International Journal of Computer Applications Recent Trends in Future Prospective in Engineering & Management Technology, pp. 20-24, ISSN : 0975-887, 2016.

21.M. Kavitha, S. Selvi, “Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons”, International Journal of Computer Science Trends and Technology, pp. 161-164, ISSN : 2347-8578, 2016.

### Author Profile



**Darshna Rajput** currently working as Assistant Professor at Department of Computer, DUIAS and DSIM&C Inst., Valsad, Gujarat. She worked as Software Developer at Hard N Soft Pvt. Ltd., Surat. She completed her graduation in B. Sc. Computer Science from VN SGU, Surat in 2004. She received the MCA degree in Computer Science from G. H. Patel Inst., S.P. University, V V Nagar in the year 2007. She is having more than 9 years of teaching experience. She has also written three books on the subject C, Java and PHP. Her area of interest, are Programming Languages, Web Designing, Data Mining and Warehousing, Database Management, etc. She now doing her Ph. D. at JJT University, Rajasthan.